The Dissertation Committee for Dandan Wang Certifies that this is the approved version of the following dissertation:

Comparing Latent Means Using Two Factor Scaling Methods: A Monte Carlo Study

Committee:

Tiffany Whittaker, Supervisor

S. Natasha Beretvas, Co-Supervisor

Barbara Dodd

Keenan Pituch

Matthew A. Hersh

Comparing Latent Means Using Two Factor Scaling Methods: A Monte Carlo Study

by

Dandan Wang, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2012

Dedicated To My Family

Acknowledgements

This dissertation cannot be completed without many people's help. First, I would like to extend my sincerest appreciation to my advisers, mentors and friends: Dr. Tiffany Whittaker and Dr. Tasha Beretvas. Dear Tiffany, thank you for bringing me to Austin and opening the door of a new world for me. During my graduate study, your encouragement, support and confidence in my ability helped me walk through every important point in this journey. No matter how difficult the situation was, every time I talked with you, I felt motivated and found confidence again. You are the person who brought not only knowledge but also "hope" to my life. And thanks for the great courses you taught, in which I found my research interests, a dissertation topic and a life-long career I want to pursue. I felt so blessed to be your student and teaching assistant since I have learned how to be a dedicated scholar and a caring and supportive mentor. Dear Tasha, thank you for always being there for me since the day one I attended UT. Whenever I came to you, you never hesitate to give me great suggestions. I was always motivated by your enthusiasm and high commitment to teaching, research and students' success. Without your guidance, I cannot successfully complete my prospectus or dissertation, not mentioning to present them at national meetings. In addition, I really appreciate that you always encouraged me to explore research and assistantship opportunities in and out of the department. They were all great experience, which highlighted my graduate school years.

I would also like to express my gratitude to other members of my dissertation committee: Dr. Barbara Dodd, Dr. Keenan Pituch and Dr. Matthew Hersh. Before they became my committee members, I have taken their courses for several semesters. Those excellent courses prepared and inspired me for my dissertation study. Additionally, their great questions and suggestions helped me improve my dissertation.

Last but not the least, I would like to express appreciation to my family members.

To my 92-year old grandma, who wrote me letters almost every month to encourage me till the day she cannot see clearly. Lao Lao, I graduated!

To my father, Yan Wang and my mother, Hong Zhou, who not only brought me to this world and gave me endless love but also taught me and supported me throughout my life. Since I was very young, my dad always told me "try your best on the things that you have committed and do not worry about results". Till today, he is still my life mentor, sharing with me his valuable life experience. My mom, my best girl friend, is always there to encourage me and bring me confidence and optimism. For every important decision I have made, from going abroad to attend graduate school to choosing a new major for my Ph.D. degree, she always supported me and encouraged me to bravely explore in a new world and to pursue my dreams.

To my husband, my soul mate and "study buddy", Yang Li, who has shared and will continue share the precious moments in my life. Yangyang, thanks for your help, understanding and love. Although we are in difference majors, you read the drafts of my prospectus and dissertation, and helped me prepare my presentations and oral exams. Without your help and love, I cannot complete my graduate study so smoothly. I am so

grateful that during graduate school years, we can study together at PCL library, working out together at the gym and have fun running around Town Lake at the weekends. Because of you, my journey is always filled with laughter and sunshine. Thank you and love you!

Comparing Latent Means Using Two Factor Scaling Methods: A Monte Carlo Study

Dandan Wang, Ph.D.

The University of Texas at Austin, 2012

Supervisors: Tiffany Whittaker, S. Natasha Beretvas

Social science researchers are increasingly using multi-group confirmatory factor analysis (MG-CFA) to compare different groups' latent variable means. To ensure that a MG-CFA model is identified, two approaches are commonly used to set the scale of the latent variable. The reference indicator (RI) strategy, which involves constraining one loading per factor to a value of one across groups, assumes that the RI has equal factor loadings across groups. The second approach involves constraining each factor's variance to a value of one across groups and, thus, assumes that the factor variances are equal across groups.

Latent mean differences may be tested and described using Gonzalez and Griffin's (2001) likelihood ratio test (LRT_{κ}) and Hancock's (2001) standardized latent mean difference effect size measure ($\hat{\delta}_{\kappa}$), respectively. Applied researchers using the LRT_{κ} and/or the $\hat{\delta}_{\kappa}$ when comparing groups' latent means may not explicitly test the assumptions underlying the two factor scaling methods. To date, no study has examined

the impact of violating the assumptions associated with the two scaling methods on latent mean comparisons.

The purpose of this study was to assess the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$ when violating the assumptions underlying the RI strategy and/or the factor variance scaling method. Type I error and power of the LRT_{κ} as well as relative parameter bias and parameter bias of the $\hat{\delta}_{\kappa}$ were examined when varying loading difference magnitude, factor variance ratio, factor loading pattern and sample size ratio. Rejection rates of model fit indices, including the χ^2 test, RMSEA, CFI, TLI and SRMR, under these varied conditions were also examined.

The results indicated that violating the assumptions underlying the RI strategy did not affect the LRT_{κ} or the $\hat{\delta}_{\kappa}$. However, violating the assumption underlying the factor-variance scaling method influenced Type I error rates of the LRT_{κ}, particularly in unequal sample size conditions. Results also indicated that the four factors manipulated in this study had an impact on correct model rejection rates of the model fit indices. It is hoped that this study provides useful information to researchers concerning the use of the LRT_{κ} and $\hat{\delta}_{\kappa}$ under factor scaling method assumption violations.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Structured means modeling (SMM) is a commonly used approach to detect latent mean differences across groups. Using the SMM approach, a confirmatory factor analysis (CFA) model that incorporates means is fit simultaneously to data sets of different groups. To ensure that the multi-group CFA (MG-CFA) model is identified, each latent variable must be assigned a scale. There are two commonly used factor scaling methods. One involves constraining one loading per factor to a value of one across groups. The observed indicator with its factor loadings constrained to a value of one across groups is called a reference indicator (RI). The second factor scaling method, which is described as the factor-variance scaling method, involves constraining each factor's variance to a value of one across groups. Both scaling methods require meeting certain assumptions. For instance, the RI strategy holds an assumption that the RI has invariant factor loadings across groups. The factor-variance scaling method, on the other hand, involves an assumption that the factor variances are equal across groups.

Previous studies have not devoted much attention to the assumptions underlying the two factor scaling methods. Only one simulation study (i.e., Johnson, Meade, & DuVernet, 2009) has investigated the impact of violating the assumption underlying the RI strategy on the accuracy of measurement invariance (MI) tests. Johnson et al. (2009) indicated that using RIs with non-invariant factor loadings led to less accurate invariance tests for a specific loading. To date, no study has examined the effect of constraining unequal factor loadings or unequal factor variances to a value of one across groups on the latent mean comparison.

Once a MG-CFA model is identified, latent mean differences across groups can be tested. The z test statistic is typically used to test the statistical significance of a latent mean difference estimate. However, Gonzalez and Griffin (2001) found that the z test was sensitive to the factor scaling method used and recommended against it when evaluating the statistical significance of a latent mean difference estimate. Lawrence and Hancock (1998) and Gonzalez and Griffin (2001), instead, have recommended using the likelihood ratio test, LRT_{κ}, when testing the significance of a latent mean difference estimate.

In addition to the statistical significance of a latent mean difference estimate, the practical significance of the latent mean difference across groups is commonly of interest. Hancock (2001) has suggested using the standardized latent mean difference effect size measure, $\hat{\delta}_{\kappa}$, which is calculated using the latent mean difference estimate divided by the squared root of the pooled factor variance across groups, as a measure of the practical significance of a latent mean difference estimate between two groups.

Several studies have investigated the impact of partial metric and/or partial intercept invariance (i.e., some of the factor loadings and/or some of the observed variable intercepts are non-invariant across groups) on the accuracy of MI tests and latent mean difference tests. It has been found that several factors influence the accuracy of MI tests and latent mean comparisons, such as group sample size ratio, factor loading pattern, loading difference magnitude and latent mean difference magnitude (Hancock, Lawrence, & Nevitt, 2000; Johnson et al., 2009; Kaplan & George, 1995; Meade & Lauthenschlager, 2004; Yang, 2008; Yoon & Millsap, 2007). However, most of these studies only investigated non-invariant factor loadings for items not serving as a RI and did not consider the implications of using RIs with non-invariant factor loadings. As mentioned before, Johnson et al. (2009) is the only study that has examined the effect of using

RIs with non-invariant factor loadings on the accuracy of measurement invariance tests. Still, their study is limited with respect to the type of MI assumptions that were tested. More specifically, Johnson et al. (2009) investigated the impact of using RIs with non-invariant factor loadings on the accuracy of the full metric invariance test and of a specific loading's invariance test, but they did not examine the impact on latent mean difference tests. Additionally, previous studies have not examined the effect of violating the assumption underlying the factor variance scaling method (i.e., constraining unequal factor variances to a value of one across groups). Thus, it is not clear how the LRT_{κ} and the $\hat{\delta}_{\kappa}$ would be affected if the assumptions underlying the RI strategy and/or the factor variance scaling method are violated.

The focus of the present study was to investigate the impact of violating the assumptions underlying the RI strategy and/or the factor-variance scaling methods on the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$. The performance of the LRT_{κ} was measured by its Type I error rates and power under specified conditions. The performance of the $\hat{\delta}_{\kappa}$, on the other hand, was evaluated by assessing its parameter bias and its relative parameter bias in certain conditions. Additionally, this simulation study evaluated the performance of five model fit indices, including the χ^2 test of model fit, CFI, TLI, SRMR and RMSEA, with respect to correct and incorrect model rejection rates. Several conditions, which are consistent with those investigated in previous studies, were manipulated in this study, including sample size ratio, non-invariant factor loading patterns for items not serving as a RI, loading difference magnitude and latent mean difference magnitude. The present study also extended previous research by including additional conditions, such as factor variance ratio and non-invariant factor loading pattern for the RI.

It is hoped that this simulation study provides researchers with useful information concerning the performance of the LRT_{κ} and $\hat{\delta}_{\kappa}$ under varying conditions that may be encountered when conducting applied research. In particular, examining the conditions that involve RIs with non-invariant factor loadings and/or unequal factor variances across groups allows researchers to be aware of the potential effects of violating the assumptions underlying these two factor scaling methods. In addition, this study intends to provide researchers with further information concerning the performance of the model fit indices when the estimating models are incorrectly specified because of a particular scaling method assumption violation.

Chapter 2: Literature Review

Many social science studies focus on comparing outcomes for groups categorized by observed variables such as gender, race, treatment group membership and so on. The ANOVA and MANOVA approaches are commonly used to compare group means based on observed scores. However, structural equation modeling (SEM) may be used to compare groups' *latent variable means*. SEM and, more specifically, multi-group confirmatory factor analysis (MG-CFA) can be used to compare, for example, male and female high school students' latent variable means on math anxiety.

There are several advantages to using the SEM approach. First, it can be used to assess the measurement invariance (MI) of a construct across groups. Second, it can be used to compare groups' latent variable means. Third, the SEM approach, which is based on a latent variable system, provides the capability to control for measurement error, yielding more accurate results than would the ANOVA/MANOVA approaches, which are based on an emergent variable system (Hancock, Lawrence, & Nevitt, 2000). When the SEM approach is used, confirmatory factor analysis (CFA) is commonly conducted to evaluate how well the hypothesized relationships between the observed variables and the latent variables account for the observed data (Bollen, 1989). In practice, the CFA modeling technique may be applied to both single-group and multiple-group data, and it can be used with theory testing as well as scale construction and validation (e.g., Babyak, Synder, & Yoshinobu, 1993; Manolis, Levin, & Gahlstrom, 1997).

The purpose of the present study was to investigate the impact of violating the assumptions associated with the two factor scaling methods (i.e., constraining one loading per

factor to a value of one across groups or constraining each factor's variance to a value of one across groups) on latent mean comparisons. This chapter begins with an introduction of the single-group CFA model, followed by the presentation of the single-group CFA model that incorporates a mean structure. Next, the multi-group confirmatory factor analysis (MG-CFA) model is introduced with a focus on measurement invariance (MI) testing across two groups. Finally, the MG-CFA model that incorporates a mean structure and its application in latent mean comparisons are introduced.

Single-Group CFA

Single-group CFA is used to evaluate how well a hypothesized model fits the observed data. Suppose a researcher is interested in high school students' math anxiety and proposes a model with four observed indicator variables to measure the construct of math anxiety. Figure 1 shows this single-group math anxiety model example. In this model, ξ represents a latent exogenous (independent) variable (math anxiety); x_1 through x_4 are measured endogenous (dependent) indicator variables (i.e., items 1 through 4 on a math anxiety scale); and δ_1 through δ_4 represent measurement errors, which are also exogenous (independent) variables. Additionally, λ_{11} to λ_{41}represent the factor loadings, which describe the relationship between the latent variable and each measured variable. The first subscript (i) of λ_{ij} represents the *i*th factor loading within the CFA model. The second subscript (*j*) represents the *j*th latent variable in the CFA model (Bollen, 1989). Given that only a single-factor CFA model was the focus in the current study, j was equal to a value of one in all conditions and, thus, it was dropped from model descriptions in the following sections.

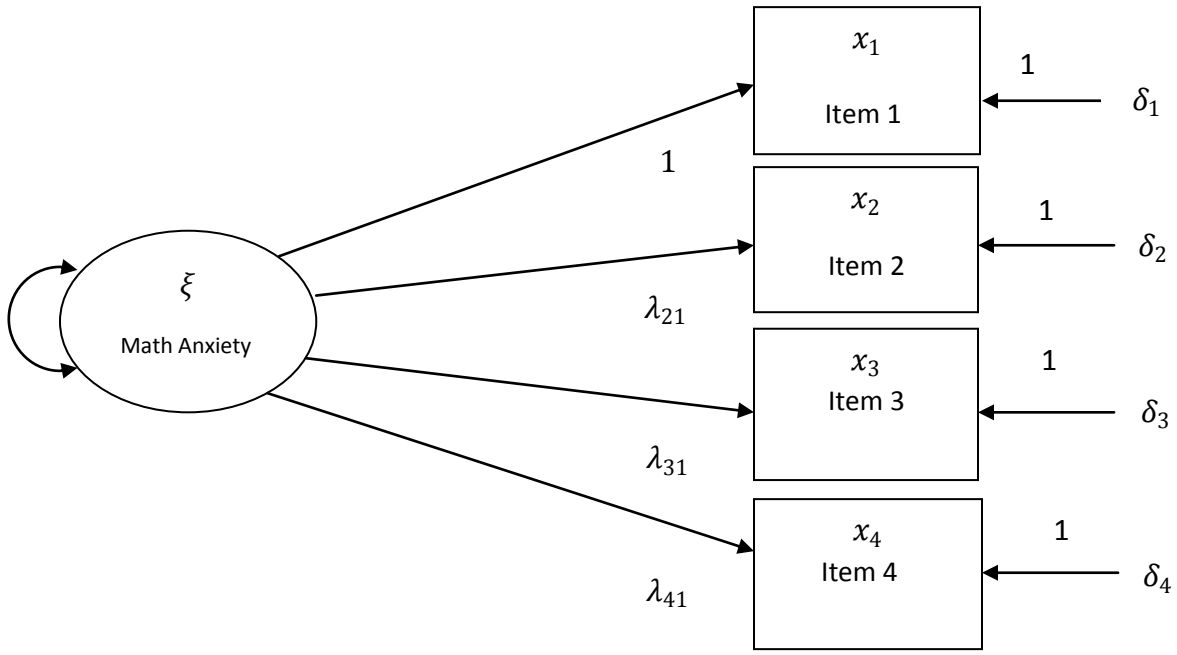

Figure 1. Single-group, single-factor confirmatory factor analysis model. ξ: a latent variable; λ_{21}- λ_{41}: factor loadings; x_1-x_4: observed indicators; δ_1- δ_4: measurement errors.

The single-group, single-factor CFA model with *p* observed indicator variables can be expressed in matrix notation using the following measurement equation:

$$\boldsymbol{x} = \boldsymbol{\Lambda\xi} + \boldsymbol{\delta}, \tag{1}$$

where $\boldsymbol{x}$ is a vector containing $\boldsymbol{p} \times \boldsymbol{1}$ observed variable scores, $\boldsymbol{\Lambda}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ vector of factor loadings which relate the observed indicator variables to the latent variable, $\boldsymbol{\xi}$ is the latent variable and $\boldsymbol{\delta}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ vector of measurement errors (Bollen, 1989). There are two important assumptions associated with Equation 1. First, the expected value of the measurement errors is equal to zero [$E(\boldsymbol{\delta}) = 0$]. Second, there is no correlation between the measurement errors and the latent variable [$E(\boldsymbol{\xi\delta}) = 0$] (Bollen, 1989). In the present study, it was also assumed that there was no correlation between the measurement errors. This assumption, however, may be relaxed.

Model Identification

Before estimation of any CFA model, the model must be identified with regard to two conditions. First, the number of unknown parameters must be less than or equal to the number of non-redundant observations in the covariance matrix of the observed variables. Second, a latent variable must have a scale of measurement (Kline, 2005). For a single-group, single-factor CFA model with *p* observed indicator variables, the number of non-redundant observations is:

$$p^* = \frac{p(p+1)}{2}, \quad (2)$$

where p represents the number of observed indicator variables and p^* is the number of non-redundant observations (Kline, 2005). Using the CFA model illustrated in Figure 1 as an example, there are $[4 \times (4 + 1)]/2 = 10$ non-redundant observations and eight unknown parameters (three factor loadings + four error variances + one factor variance) in the measurement part of the model (the reason that only three loadings are estimated is explained below). In this example, the model is over-identified (10 – 8 = 2). Thus, the fit of the model can be evaluated and the model's parameters can be estimated. When there are equal numbers of non-redundant observations and unknowns, the model is just-identified, and the proposed model will perfectly fit the data. When the number of non-redundant observations is less than the number of unknowns, the model is under-identified, and it is impossible to evaluate the model fit or estimate unknown parameters.

With regard to the second condition of model identification, two methods are commonly used to set the scale of the latent variable. One involves constraining one loading per factor to a value of one. The observed indicator variable with its factor loading constrained to a value of one is called a reference indicator (RI). This RI strategy is used to scale the latent variable in the

CFA model illustrated in Figure 1. Thus, only three of the four factor loadings are freely estimated. The second method for setting the scale of the latent variable involves constraining the factor's variance to a value of one. When this factor-variance scaling method is used, all factor loadings in the CFA model are freely estimated. For a single-group CFA model, either of these two factor scaling methods can be used to set the scale of the latent variable.

Covariance Structure Analysis

Once a CFA model is identified, how well the proposed model fits the observed data can be evaluated and unknown model parameters can be estimated. Because the input data of CFA are characteristically in the form of a covariance matrix, conducting CFA is actually testing how well the proposed model reproduces the covariances among the observed variables (Bollen, 1989). The basic hypothesis of the general CFA model is:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix implied as a function of the model parameters in $\boldsymbol{\theta}$ (Bollen, 1989). In practice, the population covariance matrix is unknown, so it is replaced by the sample covariance matrix ($\mathbf{S}$). Additionally, model parameters are unknown and they are estimated by minimizing a fitting (discrepancy) function, $F[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$. Replacing the population model parameters with the estimated model parameters, the basic equation for the CFA model becomes:

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}), \tag{4}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the implied covariance matrix and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ represents the covariance matrix implied as a function of the estimated model parameters in $\widehat{\boldsymbol{\theta}}$ (Bollen, 1989). Estimating unknown model

parameters is a process of reducing the discrepancy between each element in the sample covariance matrix and its counterpart in the implied covariance matrix. The residual matrix($\mathbf{S} - \widehat{\mathbf{\Sigma}}$) indicates how well the proposed model accounts for the observed data. The smaller the residuals, the better fit of the proposed model to the observed data (Bollen, 1989).

Maximum Likelihood Estimation

In most of the SEM software programs (e.g., AMOS, EQS, LISREL and Mplus), maximum likelihood (ML) estimation is the default estimation procedure that is used to estimate model parameters (Kline, 2005). Using this procedure, as with other SEM software estimation procedures, parameters are estimated to minimize a discrepancy function. For maximum likelihood estimation, the discrepancy function is as follows:

$$F_{ML} = ln|\mathbf{\Sigma}(\boldsymbol{\theta})| + tr[\boldsymbol{S}\mathbf{\Sigma}^{-1}(\boldsymbol{\theta})] - ln|\boldsymbol{S}| - p, \quad (5)$$

where ln is the natural log, tr is the trace function, $\mathbf{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix implied by the model parameters, $\boldsymbol{S}$ is the sample covariance matrix and p is the number of observed variables (Bollen, 1989). ML estimation maximizes the likelihood that observed data are drawn from the population. Or stated in another way, ML estimation minimizes the discrepancy between the sample covariance matrix and the model-implied covariance matrix (Bollen, 1989). The statistical assumptions associated with the ML estimation procedure in a CFA model include independence of observations, multivariate normality among observed variables and correct model specification (Kline, 2005).

The ML estimation procedure may employ full or partial information. When full-information ML estimation is conducted, all model parameters are estimated simultaneously as compared with partial-information ML estimation in which parameters are estimated in one

equation at a time. Because most of the SEM software packages employ full-information ML estimation, only this estimation procedure was used in the current simulation study.

ML estimation entails an iterative process. First, initial parameter estimate values are either defined by researchers or the default values in SEM software programs are used. Then, attempts are made to improve parameter estimates by essentially minimizing the difference between the observed and model-implied covariance matrices. The iterative process stops when the solution converges or when the maximum number of iterations has been reached. Convergence means that the improvement of the parameter estimate from one iteration to the next falls below a predetermined minimum value. For example, Mplus software implements the criterion of 0.0005. If the average improvement of the parameter estimates from one iteration to the next is equal to or smaller than 0.0005, then, convergence has been reached (Kline, 2005). For a just-identified model, the implied model will perfectly reproduce the observed covariance matrix after a few iterations. For an over-identified model, model fit will improve after iterations but the implied model will never perfectly reproduce the observed covariance matrix (Kline, 2005). If the iterations reach the pre-assigned maximum number, but the discrepancies do not fall below a predetermined minimum value, non-convergence has occurred. Because parameter estimates based on a non-convergent solution are not reliable, non-convergence should be corrected by either increasing the number of iterations or providing appropriate start values for the parameter estimates (Kline, 2005). Under some conditions, inappropriate solutions may result. For example, an estimated factor variance may be smaller than zero (negative), or a correlation between the variables may be greater than one. Such a scenario is called a Heywood case, which may be caused by model mis-specification, identification problems, inappropriate start values, or outliers (Chen, Bollen, Paxton, Curran, & Kirby, 2001). If a Heywood case is observed in the

output, researchers should search for the source of the problem, make appropriate changes and subsequently rerun the analysis (Kline, 2005). Although several other estimation procedures are available, such as unweighted least squares (ULS) and generalized least squares (GLS), the current study only involved the use of the ML estimation procedure.

Statistical Significance of the Parameter Estimate

Besides parameter estimation, the statistical significance associated with each parameter estimate is also important information and is provided by SEM software programs. The *z* test statistic, which is routinely used to evaluate the statistical significance of a parameter estimate, is calculated as:

$$z = \frac{\widehat{\boldsymbol{\theta}}}{se(\widehat{\boldsymbol{\theta}})}, \tag{6}$$

where $\widehat{\boldsymbol{\theta}}$ is the un-standardized parameter estimate of interest and $se(\widehat{\boldsymbol{\theta}})$ represents the standard error associated with the parameter estimate (Bollen, 1989).

Evaluation of Model Fit

The $\boldsymbol{\chi^2}$ test statistic. The ML estimation procedure not only provides model parameter estimates and the standard errors which allow for statistical significance tests, but it also yields model fit information to indicate how well the proposed model reproduces the observed covariance matrix. The ML-based χ^2 test statistic is the most commonly used criterion to assess model fit although other estimation procedures also provide the χ^2 statistic (Gierl & Mulvenon, 1995). The ML-based χ^2 test statistic is calculated as follows:

$$\chi^2 = (N-1)F_{ML}, \tag{7}$$

where F_{ML} is the ML estimation discrepancy function (see Equation 5) and N is the sample size (Bollen, 1989). The degrees of freedom (*df*) associated with the χ^2 test statistic are given by:

$$df_{\chi^2} = p^* - q, \tag{8}$$

where p^* is the number of non-redundant observations and q is the number of parameters to be estimated in the model (Bollen, 1989). The χ^2 test statistic is used to test the null hypothesis of model fit:

$$H_0: \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \tag{9}$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix implied as a function of population model parameters in $\boldsymbol{\theta}$. If the value of the χ^2 statistic is greater than the critical value of the χ^2 test with associated degrees of freedom, one would reject the null hypothesis of model fit. This means that the population covariance matrix is significantly different from the covariance matrix implied by the theoretical model, which indicates that the proposed model does not represent the relationships among the observed data well. On the other hand, if the value of the χ^2 test statistic is smaller than the critical value of the χ^2 test with associated degrees of freedom, one would fail to reject the null hypothesis and can infer that the proposed model fits the observed data well.

There are several limitations of the χ^2 test statistic. First, it is sensitive to sample size. Boomsma (1983) and Anderson and Gerbing (1984) indicated that the accuracy of the χ^2 estimator, $(N-1)F_{ML}$, depended on a large sample size. When the sample size is large, the value of the χ^2 statistic increases in direct proportion to $(N-1)$ (Bollen, 1989). Large χ^2 values

lead to the rejection of model fit. This means that when the sample size is large, model fit may be rejected even though the discrepancy between the population covariance matrix and the model-implied covariance matrix is negligible (Bollen, 1989; Kline, 2005). Second, the χ^2 test statistic is sensitive to the assumption of normality. Curran, West and Finch (1996) investigated the performance of the χ^2 test statistic under three distributions (i.e., a normal distribution, a moderately non-normal distribution and a severely non-normal distribution) and four sample sizes. In the estimation procedure, the model was properly specified or mis-specified. Results of their study indicated that the ML-based χ^2 test statistic was not biased under the normal distribution regardless of how the sample sizes and model specifications changed. However, it was increasingly over-estimated when non-normality (skewness and kurtoses) of the data increased. This pattern was observed for both of the properly specified and mis-specified models. Over-estimation of the ML-based χ^2 test statistic led to over-rejection of model fit, meaning that even when the proposed model accounts for the observed data well, model fit may be rejected. In sum, the χ^2 test statistic may lead to inaccurate inferences about model fit when the sample size is large or the normal distribution assumption is violated.

Supplemental model fit indices. Due to the limitations of the χ^2 test statistic, researchers have not recommended sole reliance on this single statistic, but have recommended considering supplemental model fit indices in addition to the χ^2 test statistic (Hu & Bentler, 1999; Kline, 2005; Vandenberg & Lance, 2000). Two types of model fit indices are commonly used. One is the absolute fit index, which indicates how well the proposed model accounts for the covariance among the observed variables (Hu & Bentler, 1999). Some examples of absolute fit indices include the Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI; Bentler, 1983; Jöreskog & Sörbom, 1984; Tanaka & Huba, 1985), Steiger's (1989)

Gamma Hat, McDonald's (1989) Centrality Index (Mc), the Standardized Root Mean Squared Residual (SRMR; Bentler, 1995), and the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980). Another type of fit index is the incremental fit index, which "measures the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model" (Hu & Bentler, 1999, p. 2). Some commonly used incremental fit indices include the Normed Fit Index (NFI; Bentler & Bonett, 1980), the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), the Relative Noncentrality Index (RNI; McDonald & Marsh, 1990), and the Comparative Fit Index (CFI; Bentler, 1990).

There are some conventional cutoff criteria for these various fit indices. For example, it has been recommended to use the cutoff value of 0.90 for the ML-based TLI, CFI, RNI, Gamma Hat, and Mc (Bentler, 1989; Bentler & Bonett, 1980). Any one of these model fit indices with a value greater than 0.90 indicates adequate model fit. In addition, researchers have recommended using the cutoff value of 0.05 or below to indicate acceptable fit for the SRMR and RMSEA (Browne & Cudeck, 1993; Browne & Mels, 1990; MacCallum, Browne, & Sugawara, 1996; Steiger, 1989).

Hu and Bentler (1999) evaluated the cutoff criteria for fit indices under varying conditions. In their simulation study, six sample sizes and seven conditions with respect to the distribution of the data were manipulated. In the first condition, the factors and measurement errors were normally distributed. In the second to the fourth conditions, the factors and measurement errors were not normally distributed and when uncorrelated, were independent of each other. In the last three conditions, the factors and measurement errors were not normally distributed and when uncorrelated, were dependent on each other. Based on the results of the simulation study, Hu and Bentler (1999) recommended new cutoff values for fit indices. In order

to keep the Type II error rate low and maintain the Type I error rate at an acceptable level, Hu and Bentler (1999) suggested using a value of 0.95 (equal to or greater than 0.95) for the ML-based TLI, CFI, RNI, and Gamma Hat. For the Mc, they suggested using a minimum value of 0.9 (equal to or greater than 0.9). In addition, they recommended using values equal to or less than 0.08 for the SRMR and 0.06 or below for the RMSEA to indicate acceptable model fit. Besides single cutoff values, Hu and Bentler (1999) suggested using a two-index presentation strategy. They recommended using a combination of cutoff values equal to or greater than 0.96 for the TLI (RNI, CFI, or Gamma Hat) and equal to or less than 0.09 for the SRMR, and a combination of cutoff values equal to or less than 0.09 for the SRMR and equal to or less than 0.06 for the RMSEA. They also found that, when the sample size was relatively small ($N \leq 250$), the combinations of the SRMR and the RNI, CFI, or Gamma Hat should be used to assess model fit. The combinations of these fit indices were less affected by small sample sizes than were fit index combinations that included the TLI, Mc, or RMSEA (Hu & Bentler, 1999).

Single-Group CFA with Means

The single-group CFA model discussed in the previous section focuses on the covariance among observed variables without a mean structure incorporated in the model. When only a covariance structure is analyzed, observed values are treated as mean-centered scores. Thus, the means/intercepts of observed variables are excluded from the CFA measurement equation (see Equation 1). Also, because the mean of mean-centered scores is zero, the mean of the latent variable is assumed to be zero in the covariance structure analysis (Kline, 2005). However, the mean of the latent variable and the means/intercepts of the observed variables do not have to be assumed to be equal to zero under all conditions. There are scenarios in which researchers might be interested in estimating the mean of the latent variable and the means/intercepts of the

observed variables. For example, a researcher may be interested in investigating whether female high school students' latent mean on math anxiety has changed from 11^{th} to 12^{th} grade. In such a scenario, observed variables' means/intercepts should be estimated. In addition, suppose a researcher is interested in comparing female and male high school students' latent means on math anxiety, the means of the latent variable for males and females should be estimated. Estimating latent means are more meaningful in group comparison research in the context of MG-CFA, which is introduced in the following section. In the current single-group context, a mean structure is incorporated into a CFA model along with the covariance structure to enable estimation of the latent variable mean and of observed variable means/intercepts. The discussion of the single-group CFA model with a mean structure in this section sets the stage for the following section which describes how to incorporate a mean structure into the MG-CFA model.

Figure 2 illustrates a single-group, one-factor CFA model that incorporates the estimation of the mean structure. Compared to the basic CFA model (see Figure 1), a pseudo-variable/unit predictor equal to a value of one for all individuals is added to the model and is typically represented using a triangle within which is a value of one. According to Kline (2005), this unit predictor is created automatically by SEM software programs when the analyses involve both covariance and mean structures. The latent variable and measured indicator variables are regressed on the unit predictor in the model. The unit predictor's direct effects on measured indicator variables provide the measured variables' means/intercepts (ν_x), and its direct effect on the latent variable provides the latent mean (κ) (Kline, 2005).

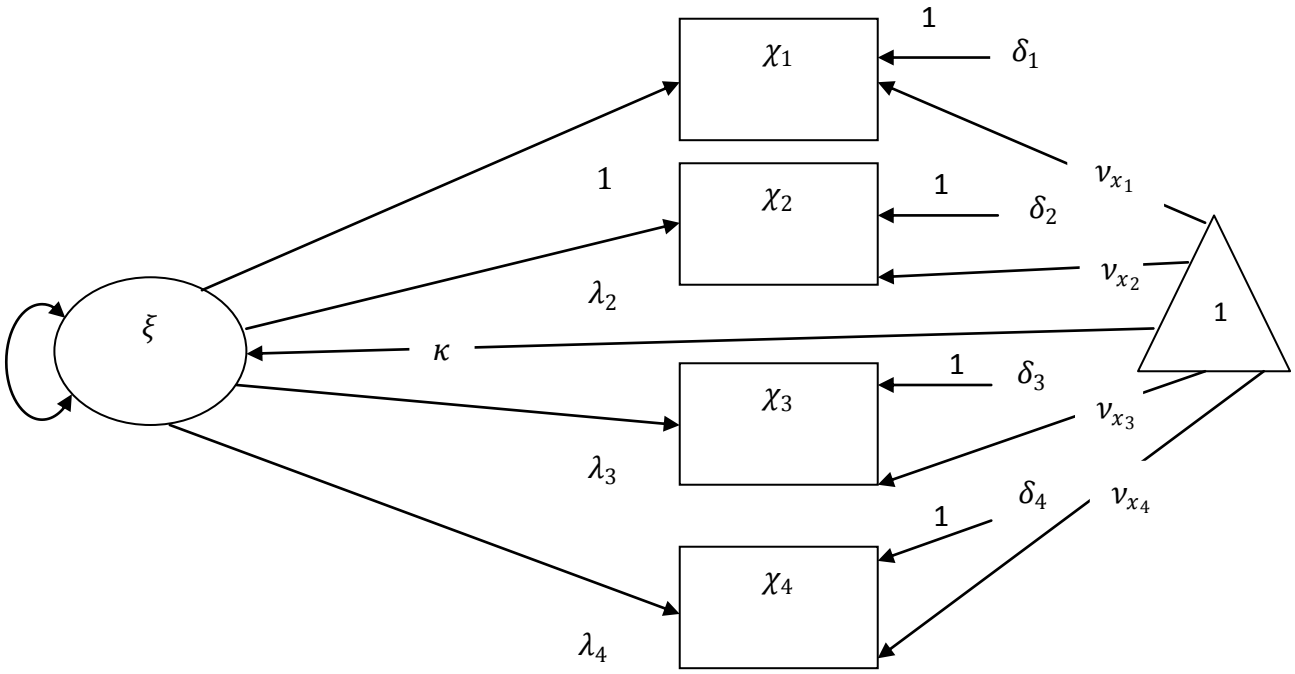

Figure 2. Single-group, single-factor confirmatory factor analysis model that incorporates means. The value of one in the triangle: a unit predictor; ν_{x_1}- ν_{x_4}: observed variable intercepts; κ: latent mean; ξ: a latent variable; λ_2- λ_4: factor loadings; x_1-x_4: observed indicators; δ_1- δ_4: measurement errors.

A CFA model with a mean structure can be expressed in a regression-type equation. For example, for a single-group, single-factor, *p*-indicator CFA model with means incorporated, the CFA measurement model can be expressed in matrix notation as:

$$\boldsymbol{x} = \boldsymbol{\nu} + \boldsymbol{\Lambda\xi} + \boldsymbol{\delta}, \tag{10}$$

where $\boldsymbol{x}$ is a vector containing $\boldsymbol{p} \times \mathbf{1}$ observed variable scores, $\boldsymbol{\nu}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of observed variables' means/intercepts, $\boldsymbol{\Lambda}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of factor loadings which relates the observed indicator variables to the latent variable, $\boldsymbol{\xi}$ represents the latent variable, and $\boldsymbol{\delta}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of measurement errors (Bollen, 1989). This equation indicates that the values of the observed variables are a function of observed variables' means/intercepts, factor loadings, the latent

variable and measurement errors. The means of $\boldsymbol{p}$ observed indicator variables (first-order moments) can be expressed in matrix notation as:

$$\boldsymbol{E}[\boldsymbol{x}] = \boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda\kappa}\,, \tag{11}$$

where $\boldsymbol{\mu}$ is a vector containing $\boldsymbol{p} \times \mathbf{1}$ expected values/means of the observed indicator variables and $\boldsymbol{\kappa}$ is the mean of the latent variable. The covariance (second-order moments) among observed indicator variables can be expressed in matrix notation as:

$$\boldsymbol{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'}] = \boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}^{'} + \boldsymbol{\Theta}, \tag{12}$$

where $\boldsymbol{\Sigma}$ represents a $\boldsymbol{p} \times \boldsymbol{p}$ covariance matrix of observed indicator variables, $\boldsymbol{\Phi}$ is the variance of the latent variable and $\boldsymbol{\Theta}$ is a $\boldsymbol{p} \times \boldsymbol{p}$ covariance matrix of measurement errors. Because the covariances among measurement errors were assumed to be zero in the current study, $\boldsymbol{\Theta}$ is simplified to a diagonal matrix containing p error variances along the diagonal of the matrix (Hancock, Lawrence, & Nevitt, 2001).

Similar to the basic CFA model, the identification of the CFA model that incorporates a mean structure also requires that the number of non-redundant observations is equal to or greater than the number of unknown parameters. For a single-group, single-factor, *p*-indicator CFA model with means incorporated, the known information includes variances of and covariances among observed indicator variables as well as the observed means of the observed indicator variables. For a single-group, single-factor, *p*-indicator CFA model with means incorporated, the number of non-redundant observations can be calculated as follows:

$$p^{**} = \frac{p(p+3)}{2}, \tag{13}$$

where p is the number of observed indicator variables and p^{**} is the number of non-redundant observations. The unknown parameters in the model include the factor variance, the factor mean, factor loadings, error variances and intercepts of the observed indicator variables. For a CFA model with a mean structure, the identification of the covariance and mean structures is considered separately. Thus, if the covariance structure is over-identified but the mean structure is under-identified, the entire model is still under-identified. Only when both the covariance and mean structures are just-identified or over-identified will unknown parameters be estimated and only when both the covariance and mean structures are over-identified is it possible to evaluate model fit (Kline, 2005). Taking the CFA model in Figure 2 as an example, there are 4(4+3)/2 = 14 non-redundant observations (4 variances, 6 covariance, and 4 observed means) and 15 unknown parameters (3 factor loadings with the loading of the reference indicator set to a value of one, 1 factor variance, 4 error variances, 4 intercepts of the observed indicator variables and 1 factor mean). In this model, the covariance structure is over-identified (10 – 8 = 2). However, the mean structure is under-identified (4 – 5 < 0) and, thus, the entire model is under-identified. It is impossible to estimate the means/intercepts of all observed indicator variables or the mean of the latent variable because estimating all these variables results in the under-identification of the entire model. In order to make the model indentified, it is necessary to impose constraints on the model. Details about the constraints that make the model identified are discussed in the latent mean comparison section.

The methods for setting the scale of the latent variable in a single-group CFA model that incorporate means are the same as those used in the basic CFA model. Thus, either one loading per factor or each factor's variance may be set to a value of one. As seen in Figure 2, one of the factor loadings is set to a value of one.

Multi-Group CFA

Single-group CFA is conducted when the purpose of the study is to evaluate how well a proposed model reproduces the data in a single group of participants. However, if researchers are interested in comparing the same CFA model across different groups, multi-group confirmatory factor analysis (MG-CFA) is the appropriate approach. Suppose a researcher is interested in assessing whether a hypothesized single-factor CFA model of high school students' math anxiety is similar for males and females, a two-group CFA model (see Figure 3) is used to evaluate model fit across the two groups. In Figure 3, the shaded box contains the grouping variable "gender". The ellipse, which is filled with dots, contains the CFA model that is tested across the two gender groups. There is an arrow pointing from gender to the ellipse, meaning that gender is the grouping variable in the two-group CFA model. The latent variable of the two-group CFA model is scaled using the RI strategy (the first item's factor loading is constrained to be equal to a value of one across groups). The asterisk (*) associated with the latent variable demonstrates that the factor variance is freely estimated across the two groups. The asterisks associated with the second through fourth factor loadings and all error variances indicate that these factor loadings and error variances are not constrained and are freely estimated across groups in this baseline two-group CFA model (Kim, Beretvas, & Sherry, 2010).

An important assumption when conducting group comparisons in the MG-CFA context is "measurement invariance" (MI), which indicates "the degree to which measurements conducted under different conditions yield psychologically equivalent measures of the same attributes" (Johnson, Meade, & DuVernet, 2009, p. 642). The MI assumption is a pre-condition for comparing the means of the latent variable across groups (Vandenberg & Lance, 2000). MI ensures that any observed differences are due to the latent variables themselves rather than due to

measurement model inequality. If the MI assumption does not hold, results of group comparisons may be inaccurate (Yang, 2008). Using the above math anxiety comparison as an example, the MI assumption ensures that the same CFA model fits the female and male data sets equally well. Thus, any differences in the observed indicators of the math anxiety factor are due to the difference in latent variable means between female and male groups rather than due to measurement model inequality. MI is oftentimes tested using the MG-CFA modeling approach. While more than two groups may be compared, only a two-group CFA model was the focus in the current simulation study.

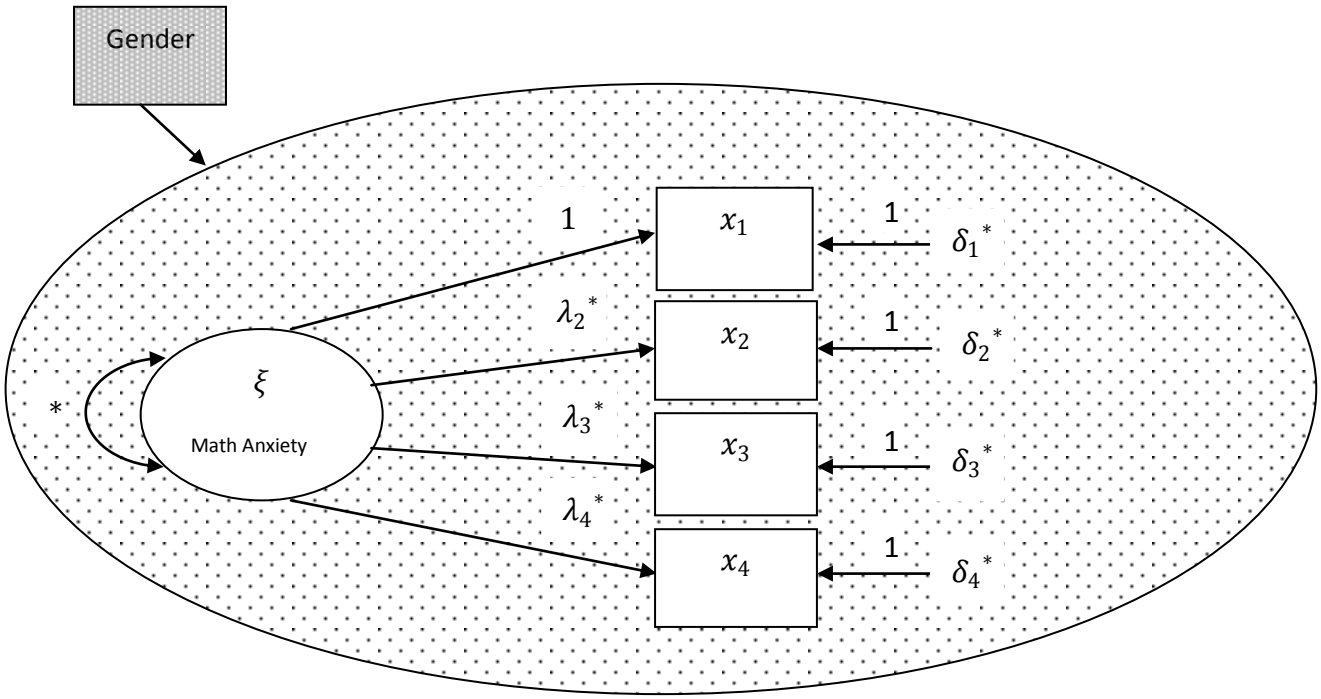

Figure 3. Two-group, single-factor, four-indicator baseline CFA model. ξ: a latent variable; λ_2-λ_4: factor loadings; x_1-x_4: observed indicators; δ_1- δ_4: measurement errors. Asterisks (*) next to the latent variable, factor loadings (besides the first one) and error variances indicate that the factor variance, factor loadings and error variances are not constrained across two gender groups.

Measurement Invariance (MI) Test

There are six possible MI tests that assess different degrees of MI assumptions, and they are generally conducted in an increasingly stringent sequence, as follows: (1) omnibus covariance equality, (2) configural invariance, (3) metric invariance, (4) scalar invariance, (5) unique variance invariance, and (6) factor variance invariance (Vandenberg & Lance, 2000). In the following section, each test is described in more detail.

Omnibus covariance equality test. The omnibus covariance equality test is typically used to test the invariance of sample covariance matrices across groups (Vandenberg & Lance, 2000). For a two-group, single-factor CFA model with *p* observed indicator variables, the null hypothesis for the omnibus covariance equality test is represented as follows:

$$H_0\colon\ \boldsymbol{\Sigma}^{\boldsymbol{g}} = \boldsymbol{\Sigma}^{\boldsymbol{g}'}, \tag{14}$$

where $\boldsymbol{\Sigma}$ represents a $\boldsymbol{p} \times \boldsymbol{p}$ population covariance matrix and the superscripts g and g' represents group membership (Vandenberg & Lance, 2000). This null hypothesis examines whether the covariance matrix for one group is equivalent to the covariance matrix for another group. In order to test the invariance of the covariance matrix across two groups, the CFA model is fitted to two groups' data simultaneously with all factor loadings (besides the factor loading of the RI), error variances and factor variances constrained to be equal across groups. The χ^2 statistic and fit indices (e.g., TLI, CFI, SRMR and RMSEA) can be used to evaluate model fit. Rejection of the null hypothesis indicates that the covariance matrix is not invariant across groups. However, this test does not indicate the sources of non-invariance. Therefore, some researchers have suggested that it is unnecessary to conduct this test (Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 1999).

Configural invariance test. The purpose of configural invariance testing is to investigate whether the same pattern of factor loadings holds across groups. This refers to the assumption that the factor model including the number of factors, the number of indicators per factor, the position of zero loadings, and the sign of non-zero factor loadings are the same across groups (Meredith, 1993; Yang, 2008). For a two-group, single-factor CFA model with *p* observed indicator variables, the null hypothesis for the configural invariance test is represented as follows:

$$H_0: \boldsymbol{\Lambda}_{\mathbf{form}}^{\boldsymbol{g}} = \boldsymbol{\Lambda}_{\mathbf{form}}^{\boldsymbol{g}'}, \tag{15}$$

where $\boldsymbol{\Lambda}_{\mathbf{form}}$ represents the pattern of factor loadings (Vandenberg & Lance, 2000). When configural invariance is tested, the CFA model without any constraints is fitted simultaneously to two groups of data and all parameters are freely estimated. This original model without any constraints is the baseline model. The χ^2 statistic and model fit indices (e.g., TLI, CFI, SRMR and RMSEA) can be used to evaluate model fit. Adequate model fit implies that the two groups employ the "same conceptual frame of reference" when responding to items on the same scale (Vandenberg & Lance, 2000, p. 37). Once evidence has been found supporting configural invariance, additional, more stringent MI assumptions can be tested (Vandenberg & Lance, 2000). However, if the null hypothesis for the configural invariance test is rejected, more stringent degrees of full MI assumptions will not be supported.

Metric invariance test. Once evidence of configural invariance has been found, metric invariance can be tested. The metric invariance test is designed to assess the invariance of corresponding items' factor loadings across groups. For a two-group, single-factor CFA model with *p* observed indicator variables, the null hypothesis for the metric invariance test is represented as follows:

$$H_0: \boldsymbol{\Lambda^g} = \boldsymbol{\Lambda^{g'}}, \tag{16}$$

where $\boldsymbol{\Lambda}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ vector of factor loadings (Vandenberg & Lance, 2000). When the metric invariance test is conducted, the CFA model is fitted simultaneously to two groups' data with all corresponding factor loadings constrained to be equal across groups. Because the model used to test metric invariance has the factor loading of each item constrained to be equal across groups, its unknown parameters are subsets of those in the baseline model in which no constraints are imposed. Thus, the model with factor loading constraints is nested within the configurally invariant baseline model. The overall difference in the two models' fit can be tested using the χ^2 difference statistic ($\Delta\chi^2$), which is calculated as follows:

$$\Delta\chi^2 = \chi^2{}_{restricted} - \chi^2{}_{baseline\ \ model}, \tag{17}$$

where $\chi^2{}_{restricted}$ is the χ^2 statistic of the model with loading constraints and $\chi^2{}_{baseline\ \ model}$ is the χ^2 statistic of the configurally invariant model. The $\Delta\chi^2$ statistic has degrees of freedom of:

$$\Delta df = df_{restricted} - df_{baseline\ \ model}, \tag{18}$$

where $df_{restricted}$ are degrees of freedom associated with the constrained model and $df_{baseline\ \ model}$ are degrees of freedom associated with the unconstrained configurally invariant model (Kline, 2005). Because the $\Delta\chi^2$ statistic asymptotically follows a non-central χ^2 distribution, its value can be compared to a critical value of the χ^2 test with associated Δdf to indicate the impact on model fit of constraining factor loadings across groups (Steiger, Shapiro, & Browne, 1985). If the $\Delta\chi^2$ value is statistically significant, the null hypothesis of metric invariance should be rejected. This means that the fit of the restricted model is significantly worse than that of the configurally invariant model, and the loading constraints should be relaxed.

In contrast, if the $\Delta\chi^2$ value is not statistically significant, metric invariance is supported, meaning that the fit of the more restricted model is comparable to that of the configurally invariant model. In this case, a more constrained form of MI can subsequently be tested. Once metric invariance has been supported, the condition of "weak measurement invariance" has been reached (Widaman & Reise, 1997).

It must be noted, however, that the $\Delta\chi^2$ statistic has the same limitations as the χ^2 test of overall model fit. That is, it is sensitive to the sample size and to violating the assumption of multivariate normality (Byrne, Shavelson, & Muthén, 1989). In addition, Yuan and Bentler (2004) found that the $\Delta\chi^2$ statistic was adversely affected by an incorrectly specified baseline model. They conducted a Monte Carlo simulation study to assess the impact of baseline model mis-specification on the $\Delta\chi^2$ statistic. Results indicated that a mis-specified baseline model led to inflated Type II errors, meaning that the $\Delta\chi^2$ test led to a failure to reject the null hypothesis too often. Thus, the presence of non-significant $\Delta\chi^2$ values did not necessarily mean that the fit of the more restricted model was comparable to that of the baseline model.

If the fit indices do not support the metric invariance model, tests may be conducted to identify which factor loadings are non-invariant. One method of identifying non-invariant factor loadings involves the use of modification indices or Lagrange Multiplier (LM) tests, which are provided by many SEM software programs (e.g., Amos, EQS, LISREL and Mplus). In the context of MG-CFA, a modification index estimates the amount by which the χ^2 statistic would decrease (the model fit would improve) if a particular cross-group constraint is relaxed (Kline, 2005). It also indicates "how poorly a particular parameter constraint is chosen" (Yoon & Millsap, 2007, p. 443). Thus, a modification index is actually the estimated χ^2 difference statistic with one degree of freedom, representing the model fit difference between the constrained model

and a model with a certain constraint released and freely estimated in both groups. Modification indices or LM tests can be used along with substantive interpretation of the validity of adding the relevant parameter to enhance model fit. In the context of MG-CFA with constraints, modification indices or LM tests can be used to help inform which parameter constraints should be released.

An alternative method of detecting which factor loadings are not invariant could involve the use of a χ^2 difference test to compare the fit of the fully metric invariant model with one in which one factor loading was allowed to vary across groups. If the χ^2 difference statistic is statistically significant, the fit of the model with the constrained loading is significantly worse than that of the model in which that factor loading is allowed to be freely estimated in both groups. This means that the relevant factor loading is not invariant across groups. This step-wise procedure could be used to identify which factor loadings are not invariant across groups.

There are different opinions about how to treat metric non-invariance. Several researchers have suggested ceasing testing of more stringent MI assumptions because the null hypothesis for testing full metric invariance has been rejected (e.g., Bollen, 1989; Millsap & Hartog, 1988). However, other researchers have recommended that partial metric invariance is an acceptable pre-condition for testing more stringent degrees of MI assumptions (Byrne et al., 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Partial metric invariance means that some, but not all, factor loadings in a MG-CFA model are invariant across groups (Kline, 2005). The recommendation of partial metric invariance is based on the belief that it represents a more realistic condition in real research (Yang, 2008). Although partial metric invariance has been discussed in several journal articles and text books (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 1999; Kline, 2005; Thompson & Green, 2006; Vandenberg & Lance, 2000),

there has been no consensus concerning the extent to which factor loading constraints may be relaxed. Steenkamp and Baumgartner (1998) have suggested that the factor loadings can be relaxed to the point that only one observed indictor in addition to the reference indicator (RI) has equivalent factor loadings across groups. In contrast, Vandenberg and Lance (2000) have recommended a more conservative partial metric invariance in which only a minority of factor loadings can be non-invariant across groups.

Scalar invariance test. Once metric invariance (or partial metric invariance) has been supported, scalar invariance can be tested. The purpose of scalar invariance testing is to assess whether observed variables' means/intercepts are invariant across groups (Yang, 2008). For a two-group, single-factor CFA model with *p* observed indicator variables, the null hypothesis for testing scalar invariance is represented as follows:

$$H_0: \boldsymbol{\tau}^{\boldsymbol{g}} = \boldsymbol{\tau}^{\boldsymbol{g}'}, \tag{19}$$

where $\boldsymbol{\tau}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of observed indicator variables' means/intercepts (Vandenberg & Lance, 2000). When scalar invariance is tested, the CFA model is fitted simultaneously to the two group's data with all factor loadings (or some of the factor loadings in partial metric invariance condition) and all intercepts constrained to be equal across groups. The null hypothesis can be tested using the χ^2 difference statistic, which is calculated by comparing the χ^2 statistic of the restricted model with that of the metric invariant (or partially metric invariant) model. Once configural, metric and scalar invariance are supported, a condition of "strong measurement invariance" has been satisfied (Millsap, 2005; Widaman & Reise, 1997). If scalar invariance does not hold as evidenced by a significant χ^2 difference statistic, additional tests may be conducted to identify non-invariant intercepts. Similar to the procedure of detecting non-

invariant factor loadings, modification indices (or LM tests) or the χ^2 difference test may be used to indentify non-invariant intercepts. As such, some of the intercept constraints can be relaxed and evidence for partial intercept invariance would be found.

According to Vandenberg and Lance (2000), the scalar invariance test is infrequently conducted among the set of MI tests. The decision of whether to conduct this test depends upon the purpose of the research. If researchers are interested in comparing latent means across groups, scalar invariance should be tested because scalar invariance as well as metric invariance ensures that any observed differences are due to the latent variable rather than due to differences in observed variables' means/intercepts or factor loadings (Hancock, 1997; Kline, 2005).

Unique variance invariance test. Once evidence for metric invariance (or partial metric invariance) has been found, unique variance invariance can be tested. The unique variance invariance test assesses the invariance of observed variables' unique (error) variances across groups. For a two-group, single-factor CFA model with *p* observed indicator variables, the null hypothesis for testing the unique variance invariance is expressed as follows:

$$H_0: \boldsymbol{\Theta}^{\boldsymbol{g}} = \boldsymbol{\Theta}^{\boldsymbol{g}'}, \tag{20}$$

where $\boldsymbol{\Theta}$ represents a $\boldsymbol{p} \times \boldsymbol{p}$ matrix with unique (error) variances of the observed indicator variables along the diagonal of the matrix and covariances among errors in the off diagonal (Vandenberg & Lance, 2000). As mentioned previously, errors were assumed to be uncorrelated in the current simulation study. Therefore, error variances, in this case, would form a diagonal matrix in which covariances among errors were all equal to zero. When unique variance invariance is tested, the model with constrained factor loadings, intercepts and unique (error) variances is simultaneously fitted to the two groups' data. The null hypothesis can be tested

using the χ^2 difference statistic, which is calculated by comparing the χ^2 value of the most restricted model with that of the metric and intercept invariant(or partially metric and intercept invariant) model. Once evidence for configural, metric, intercept and unique variance invariance has been found, the condition of "strict measurement invariance" has been satisfied (Millsap, 2005; Widaman & Reise, 1997). Some researchers have suggested that the condition of strict measurement invariance should be satisfied before latent mean comparisons can be conducted (DeShon, 2004; Millsap & Kwok, 2004). Strict measurement invariance ensures that any observed differences are due to latent mean differences across groups rather than due to covariance or mean structure inequality. However, other researchers have argued that strict measurement invariance is too restrictive to be satisfied in real research (Byrne et al., 1989; Kline, 2005; Vandenberg & Lance, 2000). According to Vandenberg and Lance (2000), unique variance invariance is less frequently tested than metric invariance and configural invariance. But it is more frequently tested than intercept invariance and latent mean differences.

Factor variance invariance test. The test of factor variance invariance assesses the equality of the factor variance across groups. For a two-group, one-factor CFA model, the null hypothesis for testing the factor variance invariance is as follows:

$$H_0: \boldsymbol{\Phi}^{\boldsymbol{g}} = \boldsymbol{\Phi}^{\boldsymbol{g}'}, \tag{21}$$

where $\boldsymbol{\Phi}$ is a $\mathbf{1} \times \mathbf{1}$ matrix containing the factor variance (Vandenberg & Lance, 2000). One important implication of rejecting this null hypothesis is that it is not appropriate to use the factor scaling method of constraining the factor variance to a value of one across groups because this method involves the assumption that factor variances are invariant across groups.

These six tests assess different degrees of MI assumptions and are typically conducted in a sequence. For example, the test of configural invariance should be conducted before the metric invariance test, and the assumption of metric invariance (at least partial metric invariance) should be supported before the scalar invariance test can be conducted. However, not all of these six tests are conducted in every study. Depending on the purpose of the research, some of the MI tests may be omitted. When a study focuses on testing the invariance of the measurement model across groups, configural invariance and metric invariance should be tested, and it is unnecessary to test scalar invariance. For example, when investigating the generalizability of the Hope Scale across genders, Babyak, Snyder and Yoshinobu (1993) tested configural invariance, metric invariance and unique variance invariance. When the purpose of the research is to compare latent means across groups, the scalar invariance test is recommended. For example, in Thompson and Green's (2006) chapter, they illustrated an example of comparing latent means across the two groups. They began with testing configural invariance and then tested metric invariance. When the null hypothesis for testing metric invariance (see Equation 16) was rejected, the authors conducted tests to isolate the sources of metric non-invariance. Under the condition of partial metric invariance, they tested intercept invariance and then, conducted tests to detect non-invariant intercepts. Finally, latent means were compared across groups under the condition of partial metric and partial intercept invariance. In sum, configural invariance and metric invariance (or partial metric invariance) should be tested and supported when testing MI (Byrne et al., 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Configural, metric and scalar invariance (or at least partial metric and scalar invariance) should be tested and supported before latent mean comparisons are conducted (Byrne, Shavelson, & Muthén, 1989; Kline, 2005).

Model Identification

The conditions required for identification of a MG-CFA model are the same as those required for a single-group CFA model in that the number of non-redundant observations is equal to or greater than the number of unknown parameters and the latent variable must have a scale of measurement. Both of the factor scaling methods (constraining one loading per factor to a value of one across groups or assigning a value of one to each factor's variance across groups) can be used to set the scale of the latent variable in a MG-CFA model. However, these two factor scaling methods are not without assumptions. The RI strategy involves the assumption that the item that has been selected to serve as a RI has invariant factor loadings across groups. The factor-variance scaling method holds an assumption that the factor variance is invariant across groups.

Latent Mean Comparison

In group comparison studies, researchers may be interested in not only the invariance of the measurement model across groups but also in latent mean differences across groups. For example, a researcher may be interested in assessing the difference between female and male high school students' latent means on math anxiety. To test latent mean differences, two SEM approaches are commonly used. One is the multiple-indicator multiple-cause (MIMIC) modeling approach and the other is the structured means modeling (SMM) approach. While both approaches can be used to compare latent means across groups, the multi-group comparison method (SMM) was the focus in the current simulation study. Discussion follows about the details of and the distinctions between these two approaches.

The MIMIC Approach

Figure 4 shows a single-factor, four-indicator MIMIC model, which is similar to the basic single-group CFA model except that a grouping variable, X_1, is included as a predictor of the latent variable in the model. The grouping variable (X_1) is modeled as having a direct effect on the latent variable. Thus, the latent variable is endogenous in this model, and the variance of its disturbance (error), ζ, is estimated. For two-group comparisons, X_1 is a dummy coded variable with zero representing the reference group and one representing the comparison group.

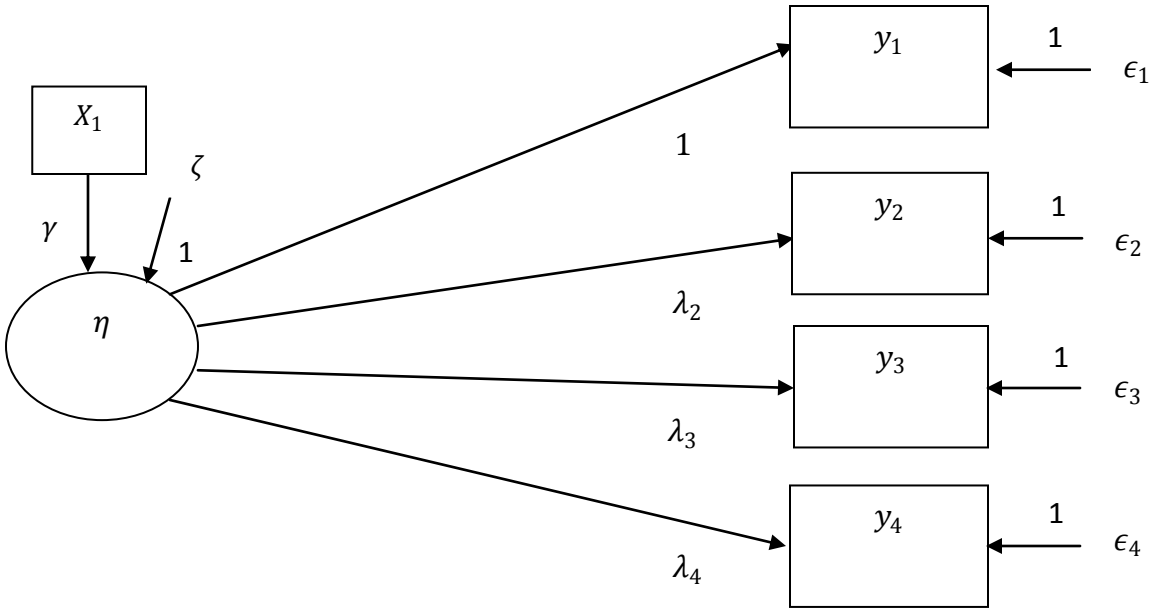

Figure 4. Multiple-indicator multiple-cause (MIMIC) model. η: an endogenous latent variable; λ_2- λ_4: factor loadings; y_1-y_4: observed indicators; ϵ_1- ϵ_4: measurement errors; X_1: a grouping variable; ζ: the error variance of the latent variable; γ: an estimate of the latent mean difference.

A single-factor, *p*-indicator MIMIC model can be expressed in matrix notation using the following measurement equation:

$$\boldsymbol{y} = \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon}, \tag{22}$$

where $\boldsymbol{y}$ is a vector containing $\boldsymbol{p} \times \boldsymbol{1}$ observed variable scores, $\boldsymbol{\Lambda}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ vector of factor loadings, $\boldsymbol{\eta}$ is the latent variable, and $\boldsymbol{\varepsilon}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ vector of measurement errors (Hancock, 1997). Because the grouping variable, X_1, has a direct effect on the latent variable in the MIMIC model, the latent variable, $\boldsymbol{\eta}$, can be expressed as:

$$\boldsymbol{\eta} = \boldsymbol{\gamma X_1} + \boldsymbol{\zeta}, \tag{23}$$

where $\boldsymbol{\gamma}$ represents the direct effect of the grouping variable, $\boldsymbol{X_1}$, on the latent variable, $\boldsymbol{\eta}$, and provides the estimate of the difference between the two groups' latent variable means. $\boldsymbol{\zeta}$ is the disturbance (error) of the latent variable (Hancock, 1997). When more than two groups are compared, more than one grouping variable is included. For example, if the MIMIC approach is used to compare latent means across three groups, two grouping variables (X_1 and X_2) are included in the model and modeled to co-vary. X_1 is a dummy coded variable such that zero represents the reference group and one represents the second group. X_2 is also a dummy coded variable with zero representing the reference group and one representing the third group. The number of grouping variables is equal to the number of groups minus one (Thompson & Green, 2006). The current simulation study focused solely on two-group comparisons.

Equation 22 demonstrates that observed variable intercepts are not estimated in the MIMIC model. This is because only the covariance matrix among observed indicator variables is analyzed when using the MIMIC approach and the mean structure is not actually incorporated into the model. When analyzing only the covariance matrix, all observed variable scores are assumed to be mean-centered. Thus, observed variable intercepts are assumed to be equal to zero and omitted from the measurement equation of the MIMIC model (Hancock, 1997).

To examine whether there is a significant difference between the two groups' latent means, a standard normal *z* test statistic can be used to evaluate the significance of the coefficient γ (see Figure 4), which provides the estimate of the difference in the two groups' latent variable means. If the coefficient is positive and statistically significant, the group of interest (coded with a one) has a significantly higher latent mean than the reference group. If the sign of the coefficient is negative and statistically significant, the latent mean of the reference group (coded with a zero) is significantly higher than that of the group of interest. It is important to note that use of the MIMIC approach involves an assumption of strict measurement invariance; that is, factor loadings, error variances and the factor variance are assumed invariant across groups. It is a very strict assumption because the conditions of error variance and factor variance invariance are unlikely to be satisfied in real-world datasets (Hancock, 1997).

The SMM Approach

The SMM approach is a multiple-group approach for testing the difference in latent variable means across groups. Figure 5 shows a one-factor, four-indicator SMM model across two groups. Similar to the single-group CFA model that incorporates means (see Figure 2), a unit predictor (constant) is included in this model, and it is modeled to have direct effects on the latent variable and observed indicator variables. Its direct effect on the latent variable represents the latent variable mean and those on observed indicator variables represent observed variables' means/intercepts (Kline, 2005). Similar to the MG-CFA model illustrated in Figure 3, there is a shaded box containing the grouping variable ("gender") in Figure 5. In addition, an ellipse, which is shaded with dots, contains the single-factor, four-indicator CFA model with means incorporated. There is an arrow pointing from the grouping variable to the ellipse, indicating that the CFA model is tested across two groups with gender as the grouping variable. Different from

the MG-CFA model in Figure 3, there is no asterisk (*) next to the factor loadings in the SMM model, meaning that all factor loadings other than the first one are constrained to be equal across groups. In addition, there is no asterisk next to the observed variable intercepts in the SMM model. This means that all observed variable intercepts are also constrained to be equal across groups. The rationale for constraining the factor loadings and observed variable intercepts to be equal across group is explained in the following section. The symbol "0/*" next to the latent mean estimate, κ, indicates that the latent mean for one group is constrained to be equal to a value of zero and the latent mean for the second group is freely estimated. Last, the asterisks next to the latent variable and error variances indicate that factor variance and error variances are freely estimated across two groups.

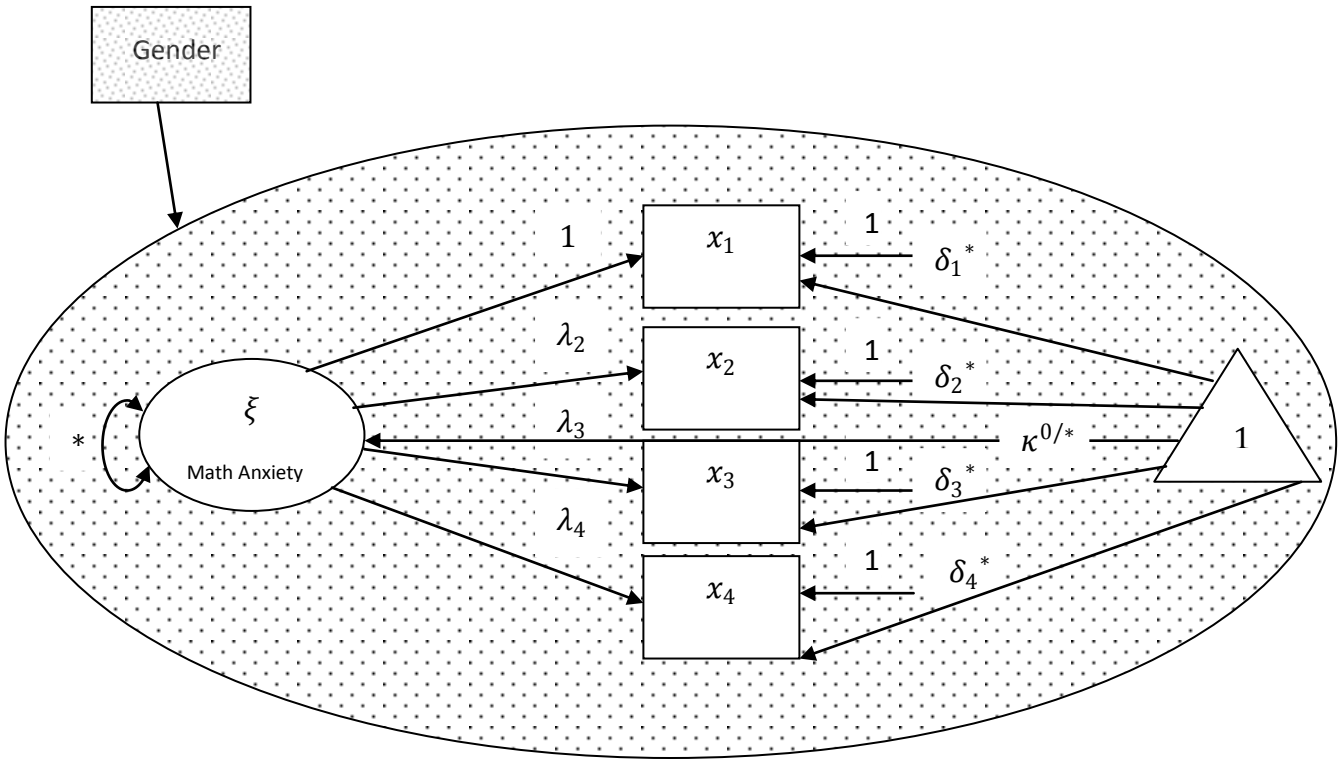

Figure 5. Structured means model (SMM). ξ: a latent variable; λ_2- λ_4: factor loadings; x_1-x_4: observed indicators; δ_1- δ_4: measurement errors. The value of one in the triangle: a unit predictor; asterisks (*) next to the latent variable and error variances indicate that the factor variance and error variances are freely estimated across groups; "0/*" next to the latent mean κ indicate that

the latent mean for group one is constrained to zero and the latent mean for group two is freely estimated.

A single-factor, *p*-indicator SMM model can be expressed in matrix notation using the following measurement equation:

$$\boldsymbol{x_g} = \boldsymbol{\nu_g} + \boldsymbol{\Lambda_g \xi_g} + \boldsymbol{\delta_g}, \tag{24}$$

where $\boldsymbol{g}$ represents group membership, $\boldsymbol{x}$ is a vector containing $\boldsymbol{p} \times \mathbf{1}$ observed variable scores, $\boldsymbol{\nu}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of indicator variable intercepts, $\boldsymbol{\Lambda}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of factor loadings which relates the observed indicator variables to the latent variable, $\boldsymbol{\xi}$ is a latent variable and $\boldsymbol{\delta}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector of measurement errors. Assuming that the mean of the measurement errors within each group is equal to zero, the expected values of observed variables in each group can be expressed in matrix notation as:

$$\boldsymbol{E}[\boldsymbol{x_g}] = \boldsymbol{\mu_g} = \boldsymbol{\nu_g} + \boldsymbol{\Lambda_g \kappa_g}\ , \tag{25}$$

where $\boldsymbol{\kappa_g}$ is the latent variable mean for group g. In addition, assuming that there is no correlation between measurement errors and the latent variable and the measurement errors are uncorrelated within each group, the covariances among observed variables in each group can be expressed in matrix notation as:

$$\boldsymbol{E}\left[(\boldsymbol{x_g} - \boldsymbol{\mu_g})(\boldsymbol{x_g} - \boldsymbol{\mu_g})^{'}\right] = \boldsymbol{\Sigma_g} = \boldsymbol{\Lambda_g \Phi_g \Lambda_g^{'}} + \boldsymbol{\Theta_g}\ , \tag{26}$$

where $\boldsymbol{\Phi}$ represents the latent variable variance and $\boldsymbol{\Theta}$ is a $\boldsymbol{p} \times \boldsymbol{p}$ diagonal matrix containing $\boldsymbol{p}$ measurement error variances (Yoon & Millsap, 2007).

As mentioned previously, parameters in a single-group CFA model with a mean structure cannot be estimated because the model's mean structure is under-identified. In a two-group CFA

model that incorporates means, the under-identification problem can be solved by a two-step strategy. First, the latent variable must be assigned a scale using either of the two factor scaling methods (constraining one loading per factor to a value of one across groups or constraining each factor's variance to a value of one across groups). Second, the reference group's latent mean is constrained to be zero and the comparison group's latent mean is freely estimated. Therefore, the test of the latent mean of the comparison group corresponds to a test of the latent mean difference between two groups because the reference group's latent mean has been constrained to be zero. To compare latent means across groups, the SMM approach generally requires that all factor loadings and observed variables' means/intercepts are invariant across groups. Therefore, corresponding loadings and intercepts are constrained to be equal across groups in model estimation procedure. If the model with all factor loadings and intercepts constrained to be equal across groups fits the observed data well, factor loadings and intercepts can be assumed invariant across group. Then, Equation 25 and Equation 26 can be respectively simplified to:

$$\boldsymbol{E}[\boldsymbol{x_g}] = \boldsymbol{\mu_g} = \boldsymbol{\nu} + \boldsymbol{\Lambda\kappa_g} \tag{27}$$

and

$$\boldsymbol{E}\left[(\boldsymbol{x_g} - \boldsymbol{\mu_g})(\boldsymbol{x_g} - \boldsymbol{\mu_g})'\right] = \boldsymbol{\Sigma_g} = \boldsymbol{\Lambda\Phi_g\Lambda}' + \boldsymbol{\Theta_g}\ , \tag{28}$$

where $\boldsymbol{\nu}$ is a $\boldsymbol{p} \times \mathbf{1}$ vector containing the observed variable intercepts which are invariant across groups, and $\boldsymbol{\Lambda}$ represents a $\boldsymbol{p} \times \mathbf{1}$ vector containing invariant factor loadings across groups (Yoon & Millsap, 2007). If the model with all factor loadings and intercepts constrained to be equal across groups does not fit the sample data well, Byrne, Shavelson and Muthén (1989) suggested that some of the factor loading constraints may be relaxed, and partial metric invariance suffices when using the SMM approach to compare latent means across groups.

Once the MG-CFA model with a mean structure is identified, the difference in two groups' latent means can be estimated and the statistical significance of the latent mean difference estimate can be tested using a standard normal *z* test statistic. If the *z* test statistic associated with the estimated latent mean difference is statistically significant, then it is inferred that the latent mean difference between two groups is statistically significant. A significant positive latent mean difference estimate, for example, indicates that the group of interest has a significantly higher latent mean than does the reference group (Hancock, 1997; Thompson & Green, 2006).

Comparing the SMM and MIMIC Approaches

The MIMIC and SMM approaches each have their own advantages. The model for the MIMIC approach is less complex than that for the SMM approach. Fewer parameters are estimated and thus a smaller sample size is needed when using the MIMIC approach (Hancock, 1997). However, the MIMIC approach involves the assumption that all factor loadings, error variances and factor variances are invariant across groups. This is a strict MI assumption that is difficult to be satisfied in practice. The SMM approach, on the other hand, permits more flexible MI assumptions; that is, some of the factor loadings constraints may be relaxed. Partial metric invariance is acceptable when using the SMM approach to conduct latent mean comparisons across groups (Byrne, Shavelson, & Muthén, 1989). Additionally, the degree of MI can be tested when utilizing the SMM approach but cannot be tested when estimating a MIMIC model. In the current study, only the SMM approach was used for comparing latent means across groups.

The z Test Statistic

When using the SMM or the MIMIC approach to test the difference in latent means across groups, the *z* test statistic has been used to evaluate the statistical significance of the latent mean difference estimate. However, the *z* test statistic has several limitations. First, it is sensitive to the format of the input data. Cudeck (1989) found that when input data were in the form of a correlation matrix instead of a covariance matrix, estimated standard errors associated with parameter estimates were "quite discrepant from the correct value" (Cudeck, 1989, p. 236). Thus, the *z* test statistics were less accurate, and may lead to incorrect inferences about the statistical significance of the parameter estimates. Second, the *z* test statistic is not invariant to the two factor scaling methods (constraining one loading per factor to a value of one across groups or constraining each factor's variance to a value of one across groups). Lawrence and Hancock (1998) and Gonzalez and Griffin (2001) indicated that when the two factor scaling methods were used, standard errors associated with parameter estimates were different, although the overall model fit was the same. More specifically, when the RI strategy was used to set the scale of the latent variable, the *z* statistic was lower than that in the condition of using the factor-variance scaling method. Consequently, as a result of the choice of method for scaling a latent variable, researchers may make different inferences about the statistical significance of the same parameter estimate.

The Likelihood Ratio Test

Because of the *z* test statistic's limitations, Hancock et al. (2000) and Gonzalez and Griffin (2001) have suggested using an alternative likelihood ratio test, LRT_{κ}, to evaluate the statistical significance of a parameter estimate in the MG-CFA model. Using the LRT_{κ}, two

models are estimated. The parameter of interest is freely estimated in one model but is constrained to be equal to zero in the second model. The χ^2 difference statistic is calculated to evaluate whether there is a statistically significant drop in model fit when constraining a particular parameter to zero. A significant χ^2 difference statistic indicates that the parameter of interest is significantly different from zero. The advantage of the LRT_κ is that it is insensitive to the factor scaling methods used. When using either of the two factor scaling methods, the χ^2 difference statistics and thus the inferences about the statistical significance of a parameter estimate are the same (Gonzalez & Griffin, 2001).

Hancock et al. (2000) used this LRT_κ to test the statistical significance of the latent mean difference estimate. Two single-factor CFA model across the two groups were estimated. In the first model, all factor loadings and observed variable intercepts were constrained to be equal across groups. In addition, the latent means of the two groups were constrained to be zero. In the second model, all factor loadings and observed variable intercepts were again constrained to be equal across groups. However, in this second model, only the reference group's latent mean was constrained to be zero and the comparison group's latent mean was freely estimated. Then, the χ^2 difference statistic was calculated by comparing the χ^2 value of the second model with that of the first model. A statistically significant χ^2 difference statistic indicated that there was a statistically significant difference between the two groups' latent means.

In sum, the LRT_κ has superiority for evaluating the statistical significance of a parameter estimate in the MG-CFA model because this technique is not sensitive to factor scaling methods. However, it must be noted that the χ^2 difference statistic also has some limitations. As mentioned in the section of MI tests, the χ^2 difference statistic is sensitive to sample size and the normality assumption. In addition, its value is affected by an incorrectly specified baseline model.

The Standardized Latent Mean Difference Effect Size Measure

Both the *z* test and the LRT_{κ} may be used to evaluate whether there is a statistically significant difference between two groups' latent means. However, they do not provide any information about the practical significance of the latent mean difference across groups. Once the null hypothesis for testing equal latent means across groups has been rejected, Hancock (2001) has suggested using a standardized latent mean difference effect size measure, $\hat{\delta}_{\kappa}$, to assess the practical difference between two groups' latent means. When the SMM approach is used to conduct latent mean comparisons across groups, the standardized latent mean difference effect size, δ_{κ}, is estimated using sample data:

$$\hat{\delta}_{\kappa} = \left|\hat{k}_1 - \hat{k}_2\right| / \hat{\phi}^{1/2}, \tag{29}$$

where $\hat{\delta}_{\kappa}$ is the estimated standardized latent mean difference effect size, $\hat{k}_1$ and $\hat{k}_2$ represent group one and group two's latent mean estimates, respectively, and $\hat{\phi}$ is the pooled factor variance estimate, which is determined as:

$$\hat{\phi} = (n_1\hat{\phi}_1 + n_2\hat{\phi}_2)/(n_1 + n_2)\ , \tag{30}$$

where $\hat{\phi}_1$ and $\hat{\phi}_2$ are the estimated factor variances for group one and group two, respectively, and n_1 and n_2 represent the sample size for group one and group two, respectively. It is important to note that the calculation and use of the pooled factor variance involves an assumption of homogeneity of the two groups' factor variances.

As mentioned before, to make the two-group CFA model with incorporated means identified, the latent mean of the reference group is typically constrained to a value of zero. Thus, Equation 29 is simplified to:

$$\hat{\delta}_{\kappa} = \left|\hat{k}_2\right| / \hat{\phi}^{1/2} , \quad (31)$$

where $\hat{k}_2$ is the latent mean estimate for the comparison group (Hancock, 1997).

The interpretation of $\hat{\delta}_{\kappa}$ is similar to that in univariate analyses in which values of 0.2, 0.5 and 0.8 represent small, moderate, and large effects, respectively (Hancock, 2001). For example, $\hat{\delta}_{\kappa} = 0.5$ can be interpreted as indicating that the two groups' latent mean estimates differ by half of a standard deviation (Hancock, 2001). However, the standardized effect size measure for latent variables is not equal to the standardized effect size measure for observed variables. Their relationship is represented as follows:

$$\hat{\delta}^{*} = \hat{\delta}_{\kappa} (H)^{1/2} , \quad (32)$$

where $\hat{\delta}^{*}$ is the standardized effect size estimate for observed variables, $\hat{\delta}_{\kappa}$ is the standardized effect size estimate for latent variables and H represents the factor's construct reliability, which is determined as:

$$H = 1/\{1 + [1/\sum_{i=1}^{p} [l_i^2/(1 - l_i^2)]]\} , \quad (33)$$

where l_i is the standardized factor loading for the ith observed indicator variable in a single-factor, p-indicator CFA model (Hancock, 2001). Construct reliability indicates the proportion of variability in a latent variable explained by the measured indicator variables in a CFA model (Hancock, 2001). Equation 32 indicates that when the factor's construct reliability is high, the value of the standardized effect size for observed variables is close to the value of standardized effect size for latent variables. Because the factor's construct reliability is rarely equal to a value of one in practice, the standard effect size measure for latent variables is typically larger than that for observed variables. Although the values of the standardized effect size measure for latent

variables and for measured variables are different, Hancock (2001) has suggested that the small difference between these two values can be ignored and that Cohen's (1988) interpretive guidelines of the standardized effect size measure, which were tested for observed variables, can be used to interpret latent variables as long as the factor's construct reliability is not too low.

In sum, Hancock's (2001) $\hat{\delta}_{\kappa}$ provides an estimate of the magnitude of the latent mean difference across groups. It is complementary to the tests which evaluate the statistical significance of a latent mean difference estimate (e.g., the *z* test and the LRT_{κ}).

Assumptions underlying the Two Factor Scaling Methods

Both of the factor scaling methods (constraining one loading per factor to a value of one across groups or assigning a value of one to each factor's variance across groups) can be used to scale the latent variable of a MG-CFA model with incorporated means. However, researchers routinely use the reference indicator (RI) strategy (e.g., Riordan & Vandenberg, 1994; Smith, Tisak, Bauman, & Green, 1991; Van de Vijver & Harsveld, 1994) because the factor-variance-based scaling method involves a strict assumption; that is, the factor variance is assumed invariant across groups. If the factor variances are not equal across groups, the assumption is violated. Then, the scale of the factor loadings will be changed, possibly making truly invariant factor loadings falsely appear non-invariant across groups. This could also make the metric invariance test less accurate (Cheung & Rensvold, 1999; Kline, 2005; Yoon & Millsap, 2007). Thus, the factor-variance scaling method is less frequently used than the RI strategy.

The reference indicator (RI) strategy is the preferred factor scaling method in the context of MG-CFA. However, this method is based on the assumption that the item which is selected to serve as a RI has equal factor loadings across groups. If this assumption is violated, all other

factor loadings in a MG-CFA model will be rescaled based on different values. For example, in a two-group, single-factor, four-indicator CFA model, the factor loadings of the two groups might be 0.8, 0.6, 0.6, 0.6 and 0.5, 0.6, 0.6, 0.6, respectively. Assume that the first item has been selected to serve as the RI in the two-group CFA model. Thus, the first factor loadings in the two groups will be multiplied by 1.25 (1/0.8 = 1.25) and 2 (1/0.5 = 2), respectively. Correspondingly, all other factor loadings in the two-group CFA model will be multiplied by 1.25 and 2, respectively. Thus, the new factor loadings of the two groups will be 1, 0.75, 0.75, 0.75 and 1, 1.2, 1.2, 1.2, respectively. Comparing the new factor loadings to the original ones, it is obvious that the factor loadings of the non-RI variables, which are equal across groups before applying the RI strategy, will now be assumed unequal. Such a result is caused by non-invariant factor loadings for the RI in the two-group CFA model. Constraining the RI's non-invariant factor loadings to a value of one across groups will thus result in different metrics for the two groups' factor loadings and can lead to incorrect inference about the changed loadings' invariance. Therefore, it is important to select an item that has invariant factor loadings across groups to serve as the RI in a MG-CFA model (Johnson, Meade, & DuVernet, 2009).

Although assumptions associated with the two factor scaling methods are important, researchers have not given the issue much attention. Johnson, Meade and DuVernet (2009) conducted a literature review of the studies that involved MI tests and were published between 2005 and 2007. They found that only 17 out of 153 studies referenced Cheung and Rensvold's (1999) study in which a new technique to select invariant item sets to serve as the RI was recommended. Most of the researchers simply assumed that the selected RI variable has invariant factor loadings across groups. To date, no simulation study has investigated the impact of using

RIs with non-invariant factor loadings or constraining unequal factor variances to a value of one across groups on latent mean comparisons.

The Reference Indicator Selection Issue

To investigate the effect of violating the assumption underlying the RI strategy, Johnson, Meade and DuVernet (2009) conducted a Monte Carlo simulation study. They manipulated four conditions, including sample size, model size, loading difference magnitude for the RI and loading difference magnitude for the non-RI variables. Additionally, they set the population factor variance to a value of one for the two groups. The likelihood ratio test (LRT_{κ}) was used to test the overall null hypothesis that all factor loadings were invariant across groups. When the null hypothesis for testing full metric invariance was not supported, the LRT_{κ} was also used to test a specific loading's invariance across groups. Results of this study indicated that improperly selected RIs (that is, selecting items with non-invariant factor loadings across groups to serve as RIs) did not affect the accuracy of full metric invariance tests. Specifically, when a RI with non-invariant factor loadings across groups was used, metric non-invariance was successfully identified. In addition, when loading difference magnitude for the RI increased, the accuracy of the full metric invariance test (that is, the rate of correctly identifying metric non-invariance) increased.

This result can be explained using the relationship between the RI and all other observed indicator variables in a MG-CFA model. When constraining the factor loading of the RI to a value of one across groups, factor loadings of all other observed indicator variables will be rescaled based on the value of the RI. As demonstrated above, if the RI has unequal factor loadings across groups, the estimated factor loadings of the non-RI variables in a MG-CFA

model will reflect this difference. Thus, the metric invariance test can then successfully detect metric non-invariance (Johnson et al., 2009).

Once the null hypothesis for testing full metric invariance is rejected, the test of a specific loading's invariance can be conducted to identify the potential source of metric non-invariance. Although the RI selection is not a critical issue for testing full metric invariance in MG-CFA, Johnson et al. (2009) indicated that using a RI with non-invariant factor loadings across groups affected the accuracy of a specific loading's invariance test. Specifically, when using a RI with non-invariant factor loadings across groups, false positive rates (the rates of identifying invariant items as non-invariant) and false negative rates (the rates of identifying non-invariant items as invariant) of a specific loading's invariance test were high.

Johnson et al. (2009) also found that there was a curvilinear relationship between the loading difference magnitude for the RI and true positive rates of the invariance test for a specific loading. More specifically, when the factor loading for the RI was invariant across groups, true positive rates of a specific loading's invariance test were high. When the loading difference magnitude for the RI increased, true positive rates of a specific loading's invariance test decreased. When the magnitudes of the loading difference for the RI and non-RI variables were equal (e.g., with the first item as the RI and the factor loadings for the first and second groups as follows: 0.6, 0.6, 0.6, 0.6 and 0.85, 0.85, 0.6, 0.6), true positive rates of a specific loading's invariance test dropped to nearly zero. Additionally, results of this study demonstrated that the sample size affected the power of a specific loading's invariance test to identify non-invariant items. When the sample size was large, both true positive and false positive rates of a specific loading's invariance test increased. Finally, results showed that the model size had no impact on the power of a specific loading's invariance test. In sum, the results of this study

indicated that selecting items with non-invariant factor loadings across groups to serve as RIs had no impact on the accuracy of the full metric invariance test. However, using RIs with non-invariant factor loadings led to low accuracy of a specific loading's invariance test.

Because using RIs with non-invariant factor loadings across groups has an adverse impact on the accuracy of a specific loading's invariance test, items with invariant factor loadings across groups should be selected to serve as RIs. An item with invariant factor loading across groups can be detected by estimating an identified MG-CFA model with a specific factor loading constrained to be equal across groups. If such a model fits the observed data well, the relevant factor loading is assumed invariant across groups. However, indentifying a MG-CFA model requires that an item with truly invariant factor loading across groups is selected to serve as the RI. Then, the question returns to how to detect an item with invariant factor loading across groups (French & Finch, 2008). Cheung and Rensvold (1999) believed that a single invariant item cannot be detected. Instead, only invariant item set(s) can be detected. They proposed using a factor-ratio test to identify invariant item set(s). An invariant item set is a group of items in which every item is invariant when tested using each of all the other items as the RI (Cheung & Resvold, 1999). The factor-ratio test systematically examines whether the ratio of the factor loading of an argument item (the item of interest) to the factor loading of a RI is invariant across groups (Cheung & Rensvold, 1999). Each item in a CFA model will be tested as an argument item and also serves as a RI when another item is tested as an argument item. For a two-group, single-factor CFA model, the null hypothesis for the factor-ratio test is expressed as follows:

$$H_0: \frac{\lambda_i^g}{\lambda_{i'}^g} = \frac{\lambda_i^{g'}}{\lambda_{i'}^{g'}}, \tag{34}$$

where λ_i is the factor loading of an argument item, i, and $\lambda_{i'}$ is the factor loading of the RI, i'. The superscripts g and g' represent group membership (Cheung & Rensvold, 1999). For a two-group, single-factor, *p*-indicator CFA model, $\frac{p(p-1)}{2}$ factor-ratio tests are needed to test all combinations of factor loadings. The null hypothesis for the factor-ratio test is tested using the χ^2 difference statistic, which is calculated by comparing the χ^2 statistic of the restricted model (with an argument item's factor loading being constrained to be equal across groups) with that of the configurally invariant model, in which all factor loadings are freely estimated across groups with the exception of the RI. If one rejects the null hypothesis, it means that the argument item being tested cannot be assumed invariant across groups when the current RI is used to set the scale of the latent variable. If one fails to reject the null hypothesis, it means that the argument item can be assumed invariant when using the current RI.

The result of each factor-ratio test is usually entered into a matrix, with one row for each argument item and one column for the relevant RI. With a single-factor, four-indicator CFA model across two groups, an illustration of a possible outcome of the factor-ratio tests is presented in Table 1. The letters (e.g., A, B, C, D, E and F) in Table 1 represent the results of the factor-ratio tests. For example, letter A in Table 1 is the result of the factor-ratio test when using item x_1 as a RI and testing the invariance of an argument item x_2. The asterisk ("*") associated with the letters indicates the statistical significance of the χ^2 difference statistic, which is used to test the null hypothesis of the factor-ratio test. Using letter B* in Table 1 as an example, it indicates that item x_3 is not invariant when using item x_1 as a RI. After filling in the matrix using the results of the factor-ratio tests, the rows and columns of the matrix are swapped to "produce the largest possible closed triangular of non-significant entries below the diagonal" (Cheung & Rensvold, 1999, p. 12). Table 2 shows the subsequently swapped matrix. In Table 2, non-

significant entries A, D and E form the largest closed triangular, indicating that an invariant item set contains item x_1, x_2, and x_4. Thus, item x_3 is a non-invariant item. Any item in the invariant item set (x_1, x_2, and x_4) can be used as the RI in the two-group CFA model. In the model estimation procedure, the factor loadings of the invariant items can be constrained to be equal across groups and the factor loading of the non-invariant item can be freely estimated. In some conditions, more than one way of swapping the matrix is available. Thus, more than one set of invariant items may be detected. The choice of the invariant item set depends on the purpose of the research and the underlying theory (Cheung & Rensvold, 1999).

Table 1

Illustration of Possible Results of Factor-Ratio Tests for a Single-Factor, Four-Indicator CFA Model across Groups

	Reference Indicator			
Argument Item	x_1	x_2	x_3	x_4
x_1	-	A	B*	D
x_2	A	-	C*	E
x_3	B*	C*	-	F
x_4	D	E	F	-

Note. The asterisks (*) next to the letters indicate that the χ^2 difference statistic is statistically significant.

Table 2

Illustration of Possible Results of Factor-Ratio Tests after Swapping the Rows and Columns

	Reference Indicator			
Argument Item	x_1	x_2	x_3	x_4
x_1	-	A	D	B*
x_2	A	-	E	C*
x_3	D	E	-	F
x_4	B*	C*	F	-

Note. The asterisks (*) next to the letters indicate that the χ^2 difference statistic is statistically significant.

French and Finch (2008) conducted a Monte Carlo simulation study to evaluate the accuracy of the factor-ratio test to detect invariant item sets. Conditions manipulated in this study

included sample size, number of factors, number of observed indicators per factor and percent of non-invariant factor loadings. The false-positive rate (the rate of identifying non-invariant item sets that are truly invariant) and the true-positive rate (the rate of correctly detecting non-invariant item sets) of the factor-ratio test were analyzed (French & Finch, 2008). The results of this study indicated that the factor-ratio test controlled false positive rates well under all conditions. Only the number of observed indicators per factor tended to influence false positive rates of the factor-ratio test. More specifically, when more indicators per factor were included in the model, false positive rates decreased. Additionally, the percent of non-invariant factor loadings, sample size or number of factors did not tend to affect false positive rates of the factor-ratio test. With regard to true positive rates of the factor-ratio test, the results demonstrated that they were influenced by the number of factors and the percent of non-invariant factor loadings. True positive rates of the factor-ratio test were higher when fewer factors were included in the model, holding the number of indicators per factor constant. In addition, when there was a large percent of non-invariant factor loadings, true positive rates of the factor-ratio test were low.

In sum, the factor-ratio test performed well in identifying non-invariant item sets when a few factors and a low percent of non-invariant factor loadings were included in a MG-CFA model. However, one shortcoming of the factor-ratio test is that it is time-consuming, particularly when a large number of observed indicators are included in the model. Thus, the factor-ratio test has been rarely used in practice (Johnson et al., 2009).

To identify non-invariant items, Yoon and Millsap (2007) alternatively recommended using modification indices. In the context of MG-CFA, a modification index or a Lagrange Multiplier (LM) test assesses whether adding a path (relaxing a cross-group constraint) would significantly improve model fit (Kline, 2005). A modification index is an estimate of the drop in

the χ^2 statistic after relaxing a cross-group constraint which is associated with a one degree of freedom difference between the model with and the model without the relevant cross-group constraint (Kline, 2005). If a modification index is statistically significant, then this means that the cross-group constraint associated with the modification index should be relaxed and a potentially non-invariant item has been detected. Modification indices are provided by most SEM software programs (e.g., AMOS, EQS, LISREL, and Mplus).

Yoon and Millsap (2007) investigated the accuracy of the modification index technique for detecting non-invariant items in a MG-CFA framework. Four conditions were manipulated in their study: (1) number of observed indicators per factor, (2) sample size, (3) percent of non-invariant factor loadings, and (4) loading difference magnitude. Yoon and Millsap (2007) did not use the RI strategy or the factor-variance scaling method to set the scale of the latent variable but instead fixed one group's factor variance to a value of one and freely estimated the second group's factor variance. In addition, all factor loadings were constrained to be equal across groups. This MG-CFA model with all factor loadings constrained to be equal across groups was fitted to the simulated data with loading non-invariance. If model fit indices indicated poor fit of the fully constrained model, meaning that the null hypothesis for testing full metric invariance is not supported, then modification indices were used to detect non-invariant items. Modification indices with values greater than 3.84 (the critical χ^2 value associated with one degree of freedom) were used to pinpoint items with non-invariant factor loadings across groups. Only one loading constraint associated with the largest modification index was relaxed at each time. This procedure continued until the largest modification index was smaller than 3.84. The items detected as having non-invariant factor loadings were then compared to those known to be non-invariant to assess the accuracy of the modification index technique.

The results of this study indicated that the modification index technique was quite successful in detecting items with non-invariant factor loadings when there was a small percent of non-invariant factor loadings in the model. This technique also worked well when the loading difference magnitude and total sample size were large. However, when there was a large percent of non-invariant factor loadings, this technique did not perform well. In sum, the results of this study showed that the modification index technique could successfully identify items with non-invariant factor loadings when the conditions are ideal (i.e., a small percent of non-invariant factor loadings, a large loading difference magnitude and a large sample size). The performance of the modification index technique is similar to that of the factor-ratio test.

Impact of Partial Measurement Invariance

Besides techniques for detecting items with non-invariant factor loadings, previous studies have investigated the impact of partial measurement (metric and intercept) invariance on MI testing and latent mean difference testing. The issue of partial metric invariance is related to the issue of the RI selection. Tests of partial metric invariance involving the use of the RI strategy for scaling the latent variable assume that the RI has invariant factor loading across groups. Because of the connection between the RI selection issue and the partial measurement invariance issue, discussion follows of studies in which the effect of partial measurement invariance has been investigated.

Kaplan and George (1995) examined the impact of partial metric invariance on the power of latent mean difference testing. They conducted a population study to assess the power, using the Wald test, to detect latent mean differences in the SMM approach under a variety of conditions. The conditions manipulated in this study included the magnitude of the latent mean

difference, sample size, frequency of non-invariant factor loadings, and the number of observed indicators per factor. Because factor loadings were varied in this study, the determinant (or the generalized variance) of the covariance matrix was thus also varied. The determinant is "a unique number associated with each square matrix" (Stevens, 2002, p. 64). It indicates the overall variance that a group of variables share. In Kaplan and George's (1995) study, sample sizes were paired with generalized variances (as measured by the determinant of the covariance matrix) to create two extra conditions. In the positive condition, the group with the larger sample size was associated with the larger generalized variance. In the negative condition, the group with the larger sample size was paired with the smaller generalized variance.

The findings of this study demonstrated that under ideal conditions (a low frequency of non-invariant factor loadings and a large sample size), the power of the latent mean difference test in the SMM approach was most affected by the true latent mean difference. More specifically, when the magnitude of the latent mean difference increased, the power of the latent mean difference test increased. Results also indicated that the sample size ratio between the two groups tended to influence the power of the latent mean difference test. When the sample sizes were equal in the two groups, the power of the latent mean difference test was less affected by non-invariant factor loadings. However, when unequal sample sizes were present, the power associated with latent mean difference testing was low even though factor loading invariance held. A large drop in the power was observed when the group sample size ratio increased. This trend was observed in both positive and negative conditions. The only difference in results between the positive and negative conditions was that the positive condition always yielded higher power than did the negative condition. Finally, when more observed indicator variables per factor were included in the model, the power of the latent mean difference test increased.

In sum, this study's results showed that the magnitude of the latent mean difference and sample size ratio had profound effects on the power of the latent mean difference test in the SMM approach. The authors also recommended using an alternative MIMIC modeling approach to test latent mean differences when the sample sizes were unequal across groups. Their recommendation was based on the MIMIC approach's superiority in handling small sample sizes. However, they did not test the Type I error rate or the power of the MIMIC approach.

Hancock, Lawrence and Nevitt (2000) expanded Kaplan and George's (1995) study by investigating how partial metric invariance affected the Type I error rate and the power of the latent mean difference tests in the SMM, MIMIC and MANOVA approaches. Although the MANOVA approach was not designed to compare latent means, it was included in Hancock et al.'s (2000) study and its performance was compared with that of the SMM and MIMIC approaches to demonstrate the potential problems of using this inappropriate approach to compare latent means. The focus of the study by Hancock et al. (2000) was to compare Type I error rates and power of the latent mean difference tests in the three approaches. The findings of the study also informed how each approach performed under varying conditions, and provided guidelines about choice of the appropriate approach under different conditions. Hancock et al. (2000) conducted a Monte Carlo simulation study to assess Type I error rates and used a population analysis to examine the power of the latent mean difference tests in the three approaches. Four conditions were manipulated in this study, including latent mean difference magnitude, total sample size, group sample size ratio and factor loading pattern. With respect to the factor loading pattern, four conditions were investigated: metric invariant condition in which all factor loadings are equal within and across the two groups, metric invariant condition in which all factor loading are equal across the two groups but different within groups, metric non-

invariant condition with approximately equivalent generalized variances (as measured by the determinant of the covariance matrix) for the two groups and metric non-invariant condition with different generalized variances for the two groups.

Some general patterns were observed for Type I error rates of the latent mean difference tests in the SMM, MIMIC and MANOVA approaches. Under the metric invariant conditions, Type I error rates of the latent mean difference tests in all three approaches were well controlled regardless of how other conditions varied. When non-invariant factor loadings were present, however, Type I error rates of the latent mean difference tests in the three approaches were at the nominal level as long as the generalized variances were approximately equivalent across the two groups. Even when the generalized variances were different in the two groups, Type I error rates of the latent mean difference tests in the SMM, MIMIC and MANOVA approaches were well controlled if the sample sizes were equal between the two groups. If both of the sample size and the generalized variance were unequal between the two groups, Type I error rates of the latent mean difference tests in the three approaches varied. The SMM approach was the only one that controlled Type I error rates well under all manipulated conditions. The MIMIC approach's Type I error rates were too low under the negative condition (when small sample sizes were paired with large generalized variances) and were too high under the positive condition (when large sample sizes were paired with large generalized variances). The opposite pattern of Type I error rates were observed for the MANOVA approach.

Hancock et al. (2000) also assessed the power of the latent mean difference tests in the MIMIC, SMM and MANOVA approaches. Under the metric invariant condition, the power of the latent mean difference tests in the three approaches increased when the sample size, magnitude of the factor loadings, and magnitude of the latent mean difference increased. When

the sample size ratio between the two groups became larger, the power of the latent mean difference tests in the three approaches decreased. Comparing the power of the MIMIC, SMM and MANOVA approaches, it seems that the power of the latent mean difference test in the MIMIC approach tended to be approximately equal to or slightly higher than that of the latent mean difference test in the SMM approach. The MIMIC approach's power superiority was more obvious when the latent mean difference was relatively large. The MANOVA approach, on the other hand, always had the lowest power among the three approaches. Under the partial metric invariant condition, the pattern of the power results for the latent mean difference tests in the three approaches were similar to those observed in the metric invariant condition as long as generalized variances were approximately equal across groups. However, if generalized variances differed and sample sizes were equal, the SMM and MIMIC approaches had similar power to detect the latent mean difference whereas the MANOVA approach had the lowest power. When different generalized variances were associated with unequal sample sizes, results showed that the SMM approach had power superiority in the negative condition whereas the MIMIC approach had higher power in the positive condition.

Based on these results, Hancock et al. (2000) recommended that the choice between the SMM and MIMIC approaches depended upon the sample size ratio. When two groups had equal sample size, both approaches were acceptable. When the sample sizes for the two groups were unequal, the SMM approach was recommended although its power was slightly lower than that of the MIMIC approach in some conditions. The choice of the SMM approach was based on its flexibility in accommodating non-invariant factor loadings. Additionally, the SMM approach had satisfactory power without sacrificing the Type I error rate. In contrast, the MIMIC approach's slightly higher power had the potential cost of the Type I error inflation (Hancock et al., 2000).

This recommendation was different from that in Kaplan and George's (1995) study. Kaplan and George (1995) recommended choosing the MIMIC approach when the sample sizes were unequal between two groups. However, they did not investigate the Type I error rate or the power of the MIMIC approach and their recommendation was solely based on the advantage of the MIMIC approach in handling small sample sizes.

Meade and Lautenschlager (2004) conducted a Monte Carlo simulation study to investigate the impact of loading non-invariance on MI testing. Partial metric invariance was simulated in the data with 17% and 67% of the items having non-invariant factor loadings across groups. Three types of loading difference were included in their design, including "all lower", "mixed" and "no difference" patterns. In the "all lower" condition, all items with non-invariant factor loadings had lower loadings in group two than in group one. In the "mixed" pattern condition, half of the non-invariant factor loadings were lower in group two and another half were lower in group one. In the "no difference" pattern condition, all factor loadings were invariant across the two groups. Other conditions manipulated in this study included total sample size and number of observed indicators per factor. The impact of partial metric invariance on MI testing was assessed through the accuracy of the omnibus test of covariance matrices and with a specific factor loading's invariance test. This study's results demonstrated that MI tests were accurate when the sample size was large and non-invariant factor loadings were in the form of a "mixed" pattern. More specifically, the omnibus covariance invariance test and a specific factor loading's invariance test were successful in detecting overall non-invariance and sources of loading non-invariance, respectively. MI tests' better performance under the "mixed" pattern condition was expected because such a pattern led to more accurate parameter estimates in each group and thus more accurate MI tests than did the "all lower" pattern (Meade & Lautenschlager,

2004). Finally, results of the study indicated that when more observed indicators per factor were included in the model, the power of MI tests increased.

The above studies have only focused on the impact of partial metric invariance on MI testing or latent mean difference testing. Yang's (2008) study provided a more complete picture of how partial metric and partial intercept invariance affected MI testing and latent mean difference testing. Yang (2008) conducted a Monte Carlo simulation study to investigate four research topics: (1) how partial metric invariance affects the detection of intercept non-invariance, (2) how partial metric invariance affects latent mean comparisons, (3) how partial intercept invariance affects latent mean comparisons, and (4) how partial metric and intercept invariance together affect latent mean comparisons. Conditions manipulated in this simulation study included model size, severity (frequency and magnitude) of loading non-invariance in the baseline model, frequency of intercept non-invariance in the target model, magnitude of the latent mean difference in the target model, construct reliability and observed score variance, and the direction of the intercept and latent mean differences. Total sample size and sample size ratio between the two groups were not manipulated. A total sample size of 500 was used and each group had a sample size of 250. The researcher used the SMM approach to conduct latent mean comparisons. In addition, the LRT_{κ} was used to detect intercept non-invariance and latent mean differences. Both Type I error rates and the power of the latent mean difference test in the SMM approach were analyzed.

Results of this study indicated that Type I error rates of the latent mean difference test in the SMM approach was not severely affected by partial metric invariance. When non-invariant factor loadings were present, Type I error rates of the latent mean difference test were retained at the nominal level. However, power of the latent mean difference test in the SMM approach

varied when conditions varied. With respect to the impact of partial metric invariance on the detection of intercept non-invariance, results showed that power of the latent mean difference test in detecting intercept non-invariance was close to the value of one when the magnitude of the intercept difference between the two groups was relatively large (e.g., 0.5). When the magnitude of the intercept difference was small, power of the latent mean difference test was low in detecting intercept non-invariance and could be increased by adding more indicators per factor to the model. In addition, results of this study demonstrated that high construct reliability led to high power of the latent mean difference test in the SMM approach. Another important finding was that the severity of the loading non-invariance did not affect the power of the latent mean difference test to detect intercept non-invariance.

With respect to the impact of partial metric invariance on the detection of the latent mean difference, patterns of the power results of the latent mean difference test were similar to those observed when investigating the impact of partial metric invariance on the detection of intercept non-invariance. With respect to the impact of partial intercept invariance on the latent mean comparison, the power of the latent mean difference test in the SMM approach was high (close to 1) when there was a relatively large latent mean difference (e.g., 0.5). Additionally, when the proportion of non-invariant intercepts decreased, the power of the latent mean difference test increased. Similar patterns of the power results were observed when examining the impact of partial metric and partial intercept invariance altogether on the latent mean comparison.

Yang (2008) also investigated the performance of modification indices in identifying non-invariant intercepts. The results indicated that the modification index technique was accurate only when the magnitude of the intercept difference was large and the percent of non-invariant intercepts was small. These findings were consistent with Yoon and Millsap's (2007) results.

Statement of the Problem

Several factors have been found to affect MI testing and latent mean difference detection. These factors include sample size ratio, factor loading pattern, loading difference magnitude, and latent mean difference magnitude. The effects of each of these factors are discussed in the following section. Additionally, factor variance ratio, which has not been investigated in previous studies, was also manipulated in the current study. The importance of considering the factor variance ratio in a MG-CFA model is explained in the following section.

Sample Size Ratio

In group comparison studies, unequal group sample size situations are more often observed than equal group sample size situations (e.g., Bowden, Lange, Weiss, & Saklofske, 2008; Sabiston & Crocker, 2007). Simulation studies conducted to investigate the impact of various conditions on MI testing and latent mean comparisons have included unequal group sample size scenarios. For example, Kaplan and George (1995) found that the power of the latent mean difference test in the SMM approach was affected by the sample size ratio between groups. More specifically, when the sample sizes were unequal in the two groups, the power of the latent mean difference test was low, even when factor loading invariance held. On the other hand, when the sample sizes were equal in the two groups, the power of the latent mean difference test was high and less affected by the frequency of non-invariant factor loadings. Hancock, Lawrence and Nevitt (2000) also varied sample size ratio between groups in their examination of conditions that may affect the Type I error rate and power of the latent mean difference tests in the SMM, MIMIC and MANOVA approaches. They found that under partial metric invariant condition, Type I error rates of the latent mean difference tests in these three approaches were at the nominal level if the sample sizes were equal in the two groups. However, Type I error rates

of the latent mean difference tests when using the MIMIC and MANOVA approaches were not well controlled when unequal sample sizes were associated with unequal generalized variances (as measured by the determinant of the covariance matrix) in the two groups. Additionally, they found that power of the latent mean difference tests tended to decrease as the sample size ratio between groups increased. Such a result was observed in the metric invariant condition and in the partial metric invariant condition with approximately equivalent generalized variances across the two groups.

To date, no study has investigated the impact of the sample size ratio between groups on the performance of the likelihood ratio test (LRT_{κ}) and the standardized latent mean difference effect size measure ($\hat{\delta}_{\kappa}$), particularly when assumptions underlying the two factor scaling methods are violated. Since it has been shown that the sample size ratio between groups affects latent mean difference detection, varied sample size ratios were included in the current simulation study to investigate their effect on the $\hat{\delta}_{\kappa}$ and the LRT_{κ}.

Factor Loading Pattern

In the current study, the factor loading pattern referred to the percent of non-invariant factor loadings and the position of the higher factor loadings in the two groups. The percent of non-invariant factor loadings has been investigated in previous simulation studies. For example, Johnson et al. (2009) varied the percent of non-invariant factor loadings when assessing how the RI and non-RI variables with non-invariant factor loadings affected the power of metric invariance testing. They found that under a large percent of non-invariant factor loadings, the power of the full metric invariance test to detect metric non-invariance was high. In contrast, Yang (2008) found that the percent of non-invariant factor loadings had no impact on the Type I

error rate or the power of the intercept invariance test. Yang (2008) also indicated that the percent of non-invariant factor loadings did not affect the Type I error rate or the power of the latent mean difference test in the SMM approach. In sum, previous research demonstrated that the percent of non-invariant factor loadings affected metric invariance tests but had no impact on intercept invariance tests or latent mean difference tests.

Although previous simulation studies have considered the percent of non-invariant factor loadings, most of them investigated non-invariant factor loadings for the non-RI variables in MG-CFA models. Only Johnson et al. (2009) investigated non-invariant factor loadings for the RI and found that using RIs with non-invariant factor loadings led to poor accuracy of a specific loading's invariance test. Due to the importance of the RI, the current study extended previous research by examining how the percent of non-invariant factor loadings for the RI and non-RI variables affected the performance of the $\hat{\delta}_{\kappa}$ and the LRT_{κ}.

Besides the percent of non-invariant factor loadings, factor loading pattern in the present study also involved the position of the higher factor loadings in the two groups. When non-invariant factor loadings are present in a two-group CFA model, two factor loading patterns may be observed. One is the "all lower" pattern in which all non-invariant factor loadings are lower in one group. The second pattern is described as the "mixed" pattern in which half of the non-invariant factor loadings favor one group and the rest favor the second group. These two patterns have been examined in previous studies. For example, Meade and Lauthenschlager (2004) included "all lower" and "mixed" patterns in their simulation design when assessing the effect of partial metric invariance on MI testing. They found that the power of MI testing was higher in the "mixed" pattern condition than in the "all lower" pattern condition. Yoon and Millsap (2007) also considered the "all lower" and "mixed" patterns when investigating the performance of the

modification index technique for detecting non-invariant items. They found that the perfect recovery rate (i.e., the rate of correctly detecting all non-invariant items) of the modification index was higher under the "mixed" pattern condition than under the "all lower" pattern condition. In addition, the "mixed" pattern increased the true detection rate of and decreased the false detection rate of the modification index technique when there was a large percent of non-invariant factor loadings.

These findings under the "mixed" pattern condition could be explained by the influence of the factor's construct reliability in confirmatory factor analysis. Construct reliability indicates "the extent to which the latent construct is reproducible from its own measured indicators" (Gagné & Hancock, 2006, p. 68). The value of construct reliability is influenced by the magnitude of the standardized factor loadings and the number of observed indicators per factor (see Equation 33). When the non-invariant factor loadings are in a "mixed" pattern, the factor loading values of the two groups are similar. Thus, the values of the factor's construct reliability are similar in the two groups. The accuracy of the parameter estimation, which is based on the factor's construct reliability, is approximately equivalent in the two groups. Consequently, the accuracy of MI testing, which is influenced by the accuracy of parameter estimations, is more ensured. In the "all lower" pattern condition, the lower factor loadings in one group results in a factor with lower construct reliability in the respective group. Correspondingly, the accuracy of the parameter estimation is different in the two groups. Thus, MI testing is less accurate (Gagné & Hancock, 2006; Yang, 2008).

In sum, the factor loading pattern, which involves the percent of non-invariant factor loadings for the RI and non-RI variables and the position of the higher factor loadings in the two

groups, is an important factor that affects MI testing. Thus, it was manipulated in the present study to investigate its impact on the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$.

Loading Difference Magnitude

Pertinent simulation studies which assessed the effect of non-invariant factor loadings on various outcomes (invariance detection and/or latent mean difference tests) in MG-CFA and its extension (SMM) have considered the loading difference magnitude. For example, Yang (2008) found that loading difference magnitude did not affect the Type I error rate or the power of the intercept invariance test, nor did loading difference magnitude has an impact on the Type I error rate or the power of the latent mean difference test in the SMM approach. Johnson et al. (2009) also included the loading difference magnitude in their design when investigating the effect of non-invariant factor loadings for the RI and non-RI variables on MI testing. Their study's results indicated that the loading difference magnitude for the RI had a positive effect on the accuracy of the full metric invariance test. More specifically, when the loading difference magnitude for the RI increased, the accuracy of the full metric invariance test increased. In contrast, the loading difference magnitude for the RI adversely affected the accuracy of a specific loading's invariance test. When the loading difference magnitude for the RI was high, the true positive rate (i.e., the rate of correctly identifying non-invariant factor loadings) of a specific loading's invariance test was low. When the loading difference magnitude for the RI was equal to that for the non-RI variables, the true positive rate of a specific loading's invariance test was approximately zero.

The loading difference magnitude has been examined in previous studies. However, no study has assessed the performance of the LRT_{κ} or the $\hat{\delta}_{\kappa}$ under varying magnitudes of the factor

loading difference. Therefore, loading difference magnitude was manipulated in this simulation study.

Latent Mean Difference Magnitude

When investigating the impact of partial MI on latent mean group comparisons, studies have varied the magnitude of the latent mean difference between groups. Kaplan and George (1995) found that latent mean difference magnitude, not the severity of loading non-invariance, had the largest impact on the power of the latent mean difference test in the SMM approach. When the latent mean difference magnitude increased, the power of the latent mean difference test also increased. Additionally, Hancock et al. (2000) showed that a large magnitude of the latent mean difference between groups led to higher power of the SMM, MIMIC and MANOVA approaches with respect to detecting latent mean differences. Yang's (2008) findings were consistent with Kaplan and George's (1995) results. The results of these simulation studies indicate that latent mean difference magnitude affects the power of latent mean difference detection. However, it is not clear how it would impact the performance of the LRT_κ and the $\hat{\delta}_\kappa$, particularly when assumptions underlying the two factor scaling methods are violated. Thus, it is necessary to consider the magnitude of the latent mean difference between groups in the current simulation study.

Factor Variance Ratio

The factor scaling method of constraining each factor's variance to a value of one across groups involves the assumption that the factor variances are invariant across groups. However, the effect of violating this assumption, that is, constraining unequal factor variances to a value of one across groups, has not been investigated in the MI literature. Several simulation studies (e.g.,

French & Finch, 2008; Hancock et al., 2000; Meade & Lautenschlager, 2004; Yang, 2008) have simply fixed the population factor variance to a value of one in their simulation designs and did not investigate the pattern of factor variances across groups. However, factor variances are not always equal across groups in applied research. For example, Kim, Cramond and Bandalos (2006) conducted a confirmatory factor analysis to investigate whether a one- or two-factor CFA model fit the observed creativity outcome data better. Results of their study indicated that the two-factor CFA model fit the data better than the one-factor CFA model. The authors also conducted MI tests to investigate whether the two-factor CFA model fit the data for the two gender groups and three grade levels equally well. The results indicated that the two-factor CFA model was more invariant across two gender groups than across three grade levels. For the author's Innovative creativity latent variable, the values of the factor variances for kindergarten, third grade and sixth grade students were 631.0, 371.4 and 309.8, respectively. For the Adaptive creativity latent variable, the values of the factor variances for kindergarten, third grade and sixth grade students were 820.5, 393.8 and 386.0, respectively. This provides evidence supporting that factor variances are not always equal across groups in applied research. Thus, it is necessary to investigate factor variance ratio in the present study.

Summary

Previous studies have investigated the effects of partial measurement (metric and/or intercept) invariance on MI testing and latent mean difference detection under various conditions. Results have shown that sample size ratio, factor loading pattern, loading difference magnitude and latent mean difference magnitude affect Type I error rates and/or the power of MI and latent mean difference tests. However, previous simulation studies did not devote much attention to the assumption underlying the RI strategy, and no study has investigated the effect of violating the

assumption underlying the factor-variance scaling method. In the current study, the impact of violating the assumptions associated with these two factor scaling methods on the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$ were examined under various conditions of sample size ratio, factor loading pattern, loading difference magnitude, latent mean difference magnitude and factor variance ratio.

Purpose of the Study

The purpose of this Monte Carlo simulation study was to investigate the performance of the likelihood ratio test (LRT_{κ}) and the standardized latent mean difference effect size measure ($\hat{\delta}_{\kappa}$), as recommended by Gonzalez and Griffin (2001) and Hancock (2001), respectively, under various conditions when using the SMM approach. Specifically, this study focused on assessing whether violating the assumptions underlying the two factor scaling methods (that is, using RIs with non-invariant factor loadings or constraining unequal factor variances to a value of one across groups) would affect the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$ within the SMM approach. The conditions that were manipulated in the current simulation study included sample size ratio, factor loading pattern, loading difference magnitude, latent mean difference magnitude and factor variance ratio. For each generated sample of data, two factor scaling methods (constraining one loading per factor to a value of one for both groups and assigning a value of one to each factor's variance for both groups) were implemented. The performance of the LRT_{κ} was evaluated through an assessment of its Type I error rates and power under specified conditions. The performance of the $\hat{\delta}_{\kappa}$ in terms of the relative parameter bias (or parameter bias under certain conditions) was also evaluated. Additionally, the performance of model fit indices,

including the χ^2 test statistic, CFI, TLI, SRMR and RMSEA, was documented in terms of their correct and incorrect model rejection rates under specified conditions.

Chapter 3: Method

A Monte Carlo simulation study was conducted to investigate the performance of the likelihood ratio test, LRT_{κ}, and the standardized latent mean difference effect size measure, $\hat{\delta}_{\kappa}$, which had been used to test the statistical significance and to assess the magnitude of latent mean differences across groups, respectively, when the assumptions underlying the RI strategy and/or the factor-variance scaling method were violated. Specifically, this study focused on examining the effects on the estimation of the LRT_{κ} and the $\hat{\delta}_{\kappa}$ in scenarios with non-invariant factor loadings and/or unequal factor variances across groups. Several conditions were manipulated in the current study, including sample size ratio, factor loading pattern, loading difference magnitude, latent mean difference magnitude and factor variance ratio. The performance of the LRT_{κ} was evaluated by examining its Type I error rates and power under specified conditions. The performance of the $\hat{\delta}_{\kappa}$was evaluated through an assessment of its relative parameter bias and parameter bias under certain conditions. In addition, the performance of model fit indices, including the χ^2 test of model fit, CFI, TLI, RMSEA and SRMR, in terms of their correct and incorrect model rejection rates was assessed.

In this chapter, the fixed design elements, which were not be manipulated in the current study, are presented first. Then, the conditions that were manipulated are discussed. Third, the procedures for generating data and estimating models are described. Finally, the data analysis procedure is discussed.

Fixed Design Elements

In this simulation study, two groups' latent variable means were compared using the SMM approach. Thus, a two-group CFA model that incorporates means (see Figure 6) was used

in both the generating and estimating models. In Figure 6, X in the shaded box represents a grouping variable. The ellipse, which is shaded with dots, contains the CFA model that was estimated across the two groups. There is an arrow pointing from the shaded box to the ellipse, meaning that X is the grouping variable for the two-group CFA model. The RI strategy was used to set the scale of the latent variable, as illustrated in Figure 6. The factor-variance scaling method was also used to scale the latent variable in the estimating models in the present study. All factor loadings (other than the factor loading of the RI) and all observed variable intercepts were constrained to be equal across groups in the estimating models. The asterisks (*) associated with the error variances indicate that the error variances were freely estimated across groups. In addition, the "0/*" associated with the latent mean, κ, indicates that the latent mean value of group one was constrained to be equal to zero whereas the latent mean value of group two was freely estimated in the estimating models.

The model that was used for generating the data and estimating model parameters contains one latent variable and six observed indicator variables. This simple model provided a reasonable starting place for the current study given that it was one of only a few assessments of the performance of the LRT_{κ} and $\hat{\delta}_{\kappa}$ in scenarios with non-invariant factor loadings and/or unequal factor variances across groups. Future research could explore more complex models. The choice of the six observed indicator variables was based on the designs of previous simulation studies and reflected what had been found in applied research. For example, Kaplan and George (1995), Meade and Lautenschlager (2004), Yoon and Millsap (2007), French and Finch (2008), and Yang (2008) all included six observed indicator variables when manipulating the number-of-indicators-per-factor condition in their simulation designs. In addition, Hinkin (1995) reviewed organizational studies that involved scale development, and found that 72

percent of the studies had six or fewer observed indicator variables. For these reasons, a single-factor, six-indicator CFA model that incorporates means was used for both data generation and model estimation in this simulation study.

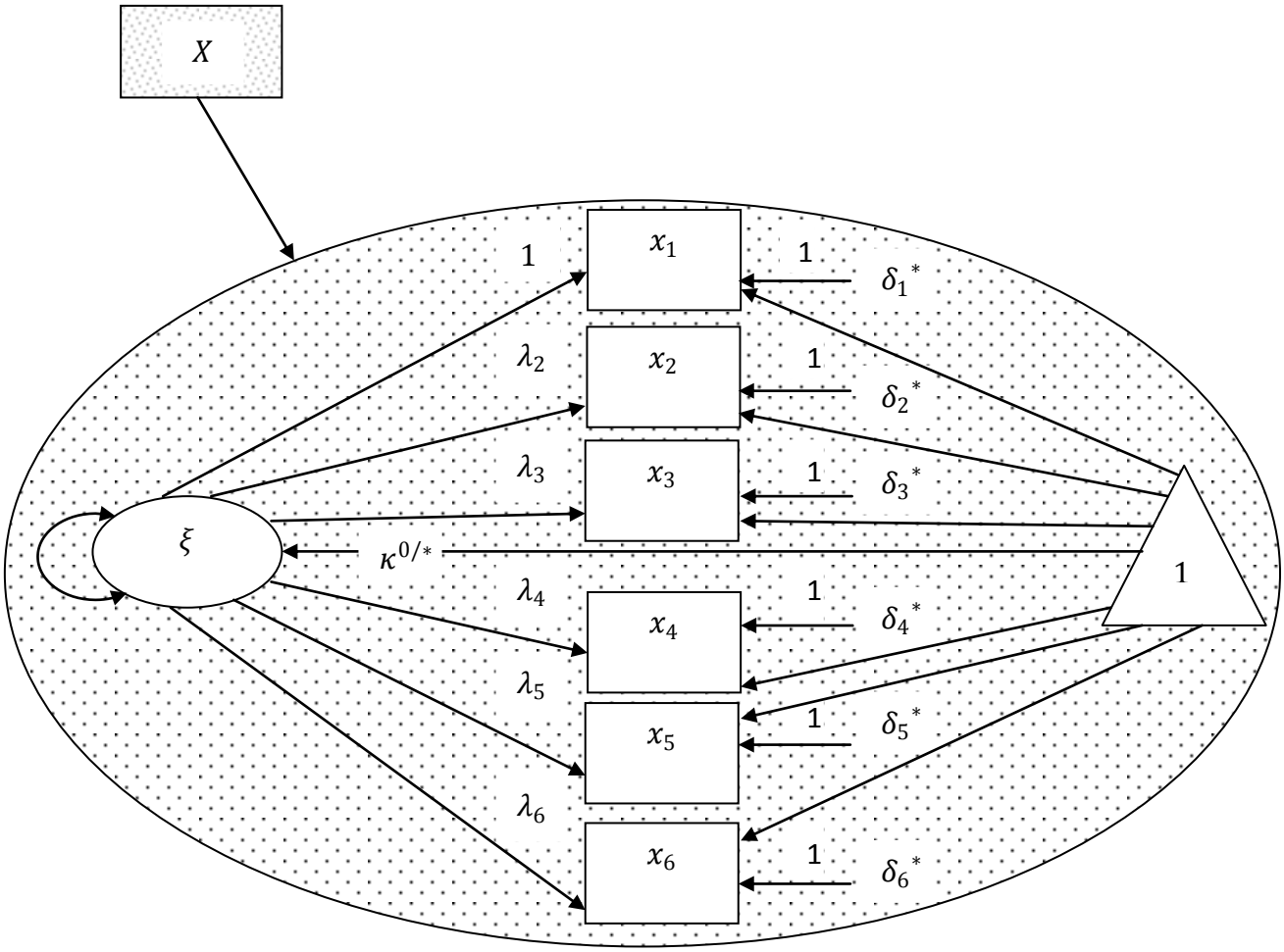

Figure 6. Two-group, single-factor, six-indicator CFA model that incorporates means. ξ: a latent variable; λ_2- λ_6: factor loadings; x_1-x_6: observed indicators; δ_1- δ_6: measurement errors; X: a grouping variable. The value of one in the triangle: a unit predictor; asterisks (*) next to the latent variable and error variances indicate that the factor variance and error variances are freely estimated across groups; "0/*" next to the latent mean κ indicates that the latent mean of group one is constrained to zero and the latent mean of group two is freely estimated.

In the generating models, the values of all invariant factor loadings were set to 0.4. Although larger factor loading values such as 0.6 and 0.7 have been used in previous simulation studies (e.g., Kaplan & George, 1995; Meade & Lautenschlager, 2004; Yang, 2008), factor

loadings of 0.4 and 0.5 have been commonly seen in applied studies (Enders & Finney, 2003). In addition, Hancock et al. (2000) included a factor loading value of 0.4 in their simulation study when comparing the performance of the SMM, MIMIC and MANOVA approaches to detect latent mean differences under a variety of conditions. Another reason for choosing the factor loading value of 0.4 for invariant factor loadings is that a relatively large loading difference value (0.4) was included in the current study (details about the loading difference magnitude is provided in the following section). In the generating models, the values of non-invariant factor loadings are equal to the value of the invariant factor loading plus the loading difference magnitude. To ensure that the values of non-invariant factor loadings are still in the range of typically seen factor loading values in applied studies, a value of 0.4 was chosen for the invariant factor loadings in the generating models.

The unique/error variance for each observed indicator variable, δ_i, in a CFA model represents the variance in the observed indicator variable that is not explained by the latent variable(s) in the model (Kline, 2005). The unique/error variance includes two parts: random variance and non-random variance. Random variance is caused by measurement unreliability. Non-random variance is the variance caused by the factor(s) other than the latent variable(s) in a CFA model (Kline, 2005). To simplify the design of the current simulation study, non-random variance was assumed to be equal to zero under all conditions (in both the generating and estimating models) and the unique/error variance, then, only includes random variance. When the factor variance(s) are set to be equal to a value of one, factor loadings in a CFA model are standardized and the unique/error variance for each observed indicator variable can be calculated using one minus the squared standardized factor loading (Meade & Lautenschlager, 2004). In the current study, conditions with truly unequal factor variances were also investigated. Under these

conditions, unique/error variances were set to be equal to their counterparts in equal-factor-variance conditions. Additionally, error covariances were set and estimated to be zero under all conditions in the current study.

For the mean structure part of the model, all observed variable intercepts were set to be equal to a value of zero across groups in the generating models because they were not the focus of the current study. The latent variable mean's value was, however, manipulated. The latent variable mean of group one was fixed at zero in the generating models and the latent variable mean of group two was set to be equal to the magnitude of the condition-specific latent mean difference, which is discussed in the following section.

Total sample size was not varied. A total sample size of 500 was used throughout. This sample size was chosen for two reasons. First, the sample size of 500 is in the range of sample sizes utilized in previous simulation research**.** For example, the sample sizes explored in Hancock et al. (2000) were 200, 400, 800 and 1,600; and in Meade and Lautenschlager (2004) were 150, 500 and 1,000. Second, Hancock et al. (2000) indicated that sufficient power was achieved with the sample size of 400 when there was a moderately large latent mean difference (0.5). Meade and Lautenschlager (2004) also found sufficient power under the sample size of 500.

Manipulated Conditions

In order to investigate the performance of the LRT_{κ} and the $\hat{\delta}_k$, several conditions were manipulated in this simulation study, including: (1) sample size ratio, (2) factor loading pattern, (3) loading difference magnitude, (4) latent mean difference magnitude, and (5) factor variance ratio. In this section, each of these conditions is described in more detail.

Sample Size Ratio

Given the previous findings concerning the effect of different group sample sizes in the related literature, three group sample size ratio conditions ($n_1 : n_2$) were used when generating the data. The equal sample size condition (1:1) was selected because it sets up a baseline condition in which the sample size in each group was 250. The two unequal sample size ratio conditions (1:4 and 4:1) were also used to generate the data. Thus, data in the 1:4 condition was generated in which the sample size was 100 and 400 in group one and in group two, respectively, and data in the 4:1 condition was generated in which the sample size was 400 and 100 in group one and in group two, respectively. These two unequal sample size conditions mimic the conditions manipulated in previous simulation research. For example, group sample size ratios of 1:1, 1:3, and 1:9 were used in Kaplan and George (1995); 1:1, 2:3, and 1:3 were used in Hancock et al. (2000); and 1:1, 1:4, and 4:1 were used in Tofighi (2005). In addition, using these two unequal sample size ratios resulted in integer values for the per-group sample sizes given the total sample size of 500.

Factor Loading Pattern

Five factor loading patterns were included in the present study. In the first or "equal" factor loading pattern condition, all factor loadings were set to be invariant across groups to serve as a baseline condition. In the second or "1st loading unequal" pattern condition, the RI's factor loading (here, the factor loading of the first observed indicator variable) was set to be non-invariant across groups. In the third or "2nd loading unequal" pattern condition, the factor loading of a non-RI variable (here, the second observed indicator variable) was set to be unequal across groups. In the second and third pattern conditions, only one factor loading was set to be non-

invariant across groups, and the relevant non-invariant factor loading had higher true value in group two than in group one by the condition-specific factor loading difference, which is subsequently described. In the fourth or "all lower" and the fifth or "mixed" pattern conditions, both the RI and the second observed indicator variable had non-invariant factor loadings across groups in the generating models. In the "all lower" pattern condition, both of the non-invariant factor loadings had lower true values (0.4) in group one. In the "mixed" pattern condition, the true factor loading value for the RI was higher in group one and the true factor loading value for the second observed indicator variable was higher in group two.

Loading Difference Magnitude

Two factor loading difference values (0.1 and 0.4) were investigated in the current simulation study. These two values are in the range of factor loading difference values investigated in previous simulation research. For example, Kaplan and George (1995) and Yang (2008) included factor loading difference values of 0.1 and 0.2 in their simulation designs. Meade and Lautenschlager (2004) used a factor loading difference value of 0.25. Hancock et al. (2000) considered factor loading difference values of 0.2 and 0.4 in their simulation research. Johnson et al. (2009) varied the loading difference magnitude from 0 to 0.4 in 0.05 increments.

Latent Mean Difference Magnitude

The current study considered two latent mean difference values (0 and 0.5). The condition of equal latent means (a zero latent mean difference) across groups was included because this permits an assessment of the Type I error rate of the LRT_{κ}. Scenarios with unequal latent means across groups were also investigated. The power of the LRT_{κ} can then be assessed. A moderately large latent mean difference value of 0.5 was included because previous simulation

studies have investigated this latent mean difference value. For example, the latent mean difference magnitude was set to 0.05, 0.15, 0.25, 0.35, and 0.5 in Kaplan and George (1995); 0.2, 0.5 and 0.8 in Hancock et al. (2000); 0, 0.2 and 0.5 in Tofighi (2005); and 0, 0.2 and 0.5 in Yang (2008). The results of these studies indicated that sufficient power was achieved when setting the latent mean difference value to 0.5. In Hancock et al.'s (2000) study, a larger latent mean difference value of 0.8 was also investigated. The results of their study indicated that the power of latent mean difference tests increased and was close to a value of one when the magnitude of the latent mean difference was increased to 0.8.

In addition to the Type I error rate and power of the LRT_{κ}, in conditions with a latent mean difference of 0 or 0.5, the performance of the $\hat{\delta}_{\kappa}$ can be assessed by evaluating its parameter bias and relative parameter bias, respectively. In sum, two latent mean difference values (0 and 0.5) were included in the current study.

Factor Variance Ratio

In this simulation study, three factor variance ratio conditions ($\Phi_1 : \Phi_2$) were considered. In the first ratio condition, the factor variances for the two groups were set to be equal (1:1). In the second and third ratio conditions, the factor variances for the two groups were set to be unequal with a ratio of 0.8:1.2 or 1.2:0.8. These two unequal factor variance conditions represent a moderate difference between the two groups' factor variances that provides a starting point for this line of research.

Study Design Overview

The conditions investigated in the current simulation study were not fully crossed. For example, in the "equal factor loading" conditions, only sample size ratio, factor variance ratio

and latent mean difference magnitude varied. The conditions that were manipulated in the present study resulted in 162 design cells, which included 18 designs cells in the equal factor loading conditions [3 (sample size ratios) x 2 (latent mean difference magnitudes) x 3 (factor variance ratios)] and 144 design cells in the unequal factor loading conditions [3 (sample size ratios) x 4 (factor loading patterns) x 2 (loading difference magnitudes) x 3 (factor variance ratios) x 2 (latent mean difference magnitudes)]. The Type I error rate and power of the LRT_{κ}, relative parameter bias and parameter bias of the $\hat{\delta}_{\kappa}$ and model rejection rates of model fit indices, including the χ^2 test of model fit, CFI, TLI, SRMR and RMSEA, were examined under specified conditions. Table 3 and Table 4 illustrate the dimensions of the study design.

Table 3

Dimensions of the Study Design

Sample Size Ratio ($n_1 : n_2$)
1 : 1
1 : 4
4 : 1
Loading Difference Magnitude
0.1
0.4
Latent Mean Difference Magnitude
0.0
0.5
Factor Variance Ratio ($\Phi_1 : \Phi_2$)
1.0 : 1.0
1.2 : 0.8
0.8 : 1.2

Table 4

Factor Loading Patterns

Factor Loading Pattern	All Factor Loadings Invariant	RI with Invariant Loadings	2nd Item with Invariant Loadings	Non-invariant loadings have lower true values in group one
All equal	Yes	Yes	Yes	-
1st loading unequal	No	No	Yes	Yes
2nd loading unequal	No	Yes	No	Yes
All lower	No	No	No	Yes
Mixed	No	No	No	No

Note. A "-"indicates that there is no non-invariant factor loading. Abbreviations used in this table are explained in Table 6.

Data Generation

Raw data for the two groups was generated using SAS 9.2 software (SAS Institute Inc., 2008). First, the specified population parameters (factor loadings, factor variances, error variances as well as latent variable means by group) were substituted into the relevant parameter equations (Equations 11 and 12) to obtain the generating covariance matrices and mean vectors. Next, the Kaiser and Dickman's (1962) matrix decomposition procedure was implemented for generating the data assuming a multivariate normal distribution with desired inter-variable relationships and means of the observed variables (Fan & Fan, 2005). Each generated sample of data consists of $n_1 \times 6$ and $n_2 \times 6$ matrices for group one and group two, respectively, where n_1 and n_2 represent the condition-specific sample size for each of the two groups. To achieve accurate results, 1,000 replications were conducted for each of the 162 combinations of conditions. Models were be fitted to each generated data set to investigate the performance of the LRT_κ, the $\hat{\delta}_\kappa$ and model fit indices under specified conditions.

Model Estimation

Once raw data for the two groups was generated, Gagné and Furlow's (2009) procedure was used in which SAS 9.2 was programmed to call DOS to run M*plus* 6.1 software (Muthén & Muthén, 2010) to estimate the models. Maximum likelihood (ML) estimation, which is the default estimation procedure in M*plus* 6.1, was used to estimate all models in the current simulation study. A single-factor, six-indicator CFA model with all factor loadings (besides the factor loading of the RI) and all observed variable intercepts constrained to be equal across groups (see Figure 6) was fitted simultaneously to each generated data set for the two groups. For each generated sample of data, two factor scaling methods were used to set the scale of the latent variable. When using the RI strategy, the first factor loading was constrained to be equal to a value of one across groups and all other factor loadings were constrained to be equal across groups. The factor variances for the two groups were freely estimated. When the factor-variance-based scaling method was implemented, the factor variance was instead constrained to be equal to a value of one across groups. In addition, all factor loadings were constrained to be equal across groups. It is important to note that using these two factor scaling methods resulted in models with different degrees of freedom. More specifically, the models using the factor-variance scaling method had one more degree of freedom than the models using the RI strategy. Thus, for the same generated data set, the model using RI strategy had a slightly better model fit than the model using the factor-variance scaling method. Because the purpose of the current study was to investigate how violating the assumptions underlying the RI strategy and/or the factor-variance scaling method would affect the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$, comparing the performance of the LRT_{κ} and the $\hat{\delta}_{\kappa}$ when using the two factor-scaling methods was not the

focus. Thus, different degrees of freedoms, which are caused by using two different factor scaling methods, did not provide any problems in this study.

In the estimating models, some cross-group constraints were imposed on observed variable intercepts and latent variable means. All observed variable intercepts were constrained to be equal across groups because their population values were all set to be equal to zero. Additionally, because the purpose of this simulation study included examining the performance of the LRT_{κ} in terms of the Type I error rate and power under specified conditions, two kinds of CFA models were tested for each generated data set. In the first kind of model, all factor loadings (besides the factor loading of the RI) and observed variable intercepts were constrained to be equal across groups. In addition, the latent mean of group one was constrained to be equal to zero and the latent mean of group two was freely estimated. In the second kind of model, factor loadings and intercepts were constrained as in the first kind of model. However, the latent means for both groups were constrained to be equal to zero. This second kind of model's estimation was only used to calculate the LRT_{κ}. Thus, for each generated data set, four models (two factor scaling methods x two ways of constraining latent means) were estimated. Last, error variances were not constrained to be equal across group in the estimating models and they were freely estimated.

Estimates of the latent mean for group two and factor variances for the two groups were saved for the model with one latent mean freely estimated (the first kind of model). These parameter estimates were used to estimate the standardized latent mean difference effect size, $\hat{\delta}_{\kappa}$, (see Equation 31). Model fit indices, including the χ^2 statistic, CFI, TLI, SRMR and RMSEA, were also kept for this model. In addition, the χ^2 statistic for the model in which both groups' latent variable means were constrained to be equal to zero (the second kind of model) was saved.

The two models' χ^2 statistics were used to calculate the χ^2 difference statistic (see Equation 17), which was used to calculate the LRT_κ to test the statistical significance of the latent mean difference estimate using the nominal alpha level of 0.05.

When estimating models, non-convergence (that the improvement of a parameter estimate does not fall below a pre-determined minimum value when the iterations have reached the pre-assigned maximum number) and/or Heywood cases (i.e., improper results, such as when a correlation value is greater than one and a variance value is smaller than zero) may be encountered. If these results were observed, the conditions and the model being estimated were recorded. In addition, new data was generated and the model estimation procedure was implemented again until 1,000 converged and proper solutions were obtained in each design cell.

Data Analysis

This section describes how the performance of the likelihood ratio test, the standardized latent mean difference effect size measure and model fit indices were evaluated.

Performance of the Likelihood Ratio Test

The performance of the LRT_κ was evaluated by summarizing its Type I error rates and power rates by condition. The Type I error rate of the LRT_κ is defined as the proportion of incorrect rejections of the null hypothesis of equal latent means, out of the 1,000 converged replications, in conditions with equal latent means across groups. Type I error rates of the LRT_κ were evaluated using Bradley's (1978) liberal criterion of $\alpha \pm 1/2\alpha$ such that if Type I error rates were less than 2.5% then they were considered overly conservative. If rates were greater than 7.5%, they were considered overly liberal. The power of the LRT_κ is defined as the proportion of correct rejections of the null hypothesis of equal latent means, out of the 1,000

converged replications, in conditions with unequal latent means across groups. There are two commonly used power criteria, 0.80 and 0.90 (Goodman & Berlin, 1994). In the current study, the more stringent power criterion 0f 0.90 was used, as recommended by Rossi (1990) and Cashen and Geiger (2004), to evaluate the performance of the LRT_{κ}.

Performance of the Standardized Latent Mean Difference Effect Size Measure

The standardized latent mean difference effect size measure, $\hat{\delta}_{\kappa}$, was introduced by Hancock (2001) to assess the practical significance of a latent mean difference estimate across groups. In this study, the performance of the of the $\hat{\delta}_{\kappa}$was examined through an assessment of its relative parameter bias and parameter bias under certain conditions. The relative parameter bias of the $\hat{\delta}_{\kappa}$ is calculated as follows:

$$RPB\left(\hat{\delta}_{\kappa}\right) = \frac{\bar{\hat{\delta}}_{\kappa} - \delta_{\kappa}}{\delta_{\kappa}}, \tag{35}$$

where δ_{κ} is the population standardized latent mean difference effect size and $\bar{\hat{\delta}}_{\kappa}$ is the mean of the estimates of δ_{κ} across the 1,000 converged replications under the conditions for which δ_{κ} is not equal to zero (Hoogland & Boomsma, 1998). In conditions in which the δ_{κ} is equal to zero, parameter bias, instead of relative parameter bias, of the $\hat{\delta}_{\kappa}$ is calculated as:

$$B\left(\hat{\delta}_{\kappa}\right) = \bar{\hat{\delta}}_{\kappa} - 0, \tag{36}$$

According to Hoogland and Boomsma (1998), conditions in which the $\left|RPB\left(\hat{\delta}_{\kappa}\right)\right|$ or $\left|B\left(\hat{\delta}_{\kappa}\right)\right|$ are less than 0.05 indicate acceptable bias in the $\hat{\delta}_{\kappa}$.

Performance of Model Fit Indices

In the present study, the estimating models were mis-specified in a subset of conditions (i.e., when factor loadings and/or factor variances were set to be unequal across groups in the generating models). The performance of model fit indices, including the ML-based χ^2 statistic, CFI, TLI, SRMR and RMSEA, was therefore evaluated by examining the correct model rejection rate that is the proportion of replications in which the model fit indices correctly reject the null hypothesis of model fit. Additionally, the estimating models were correctly specified in some conditions examined in the current study (i.e., when all factor loadings were set to be invariant across groups and factor variances were set to be equal across groups in the generating models). When the models were correctly specified, the performance of the five model fit indices was assessed using the incorrect model rejection rate, which is the proportion of replications in which the model fit indices incorrectly reject the null hypothesis of model fit. An alpha level of 0.05 was used to evaluate the statistical significance of the χ^2 statistic. Two CFI and TLI cutoff values, 0.90 and 0.95, which were proposed by Bentler and Bonnet (1980) and Hu and Bentler (1999), respectively, were used to determine whether the null hypothesis of model fit should be rejected. Two SRMR cutoff values, 0.05 and 0.08, which were suggested by Steiger (1989) and Hu and Bentler (1999), respectively, were used in the current study. Additionally, two RMSEA cutoff values, 0.05 and 0.06, which were suggested by Steiger (1989) and Hu and Bentler (1999), respectively, were used to evaluate model fit.

Chapter 4: Results

The focus of this Monte Carlo simulation study was to investigate the impact of violating the assumptions underlying the two factor scaling methods (i.e., constraining one loading per factor to a value of one across groups or constraining each factor's variance to a value of one across groups) on the testing and estimation of the latent mean difference across groups. The likelihood ratio test (LRT_κ), which had been used to test the statistical significance of a latent mean difference estimate across groups, was evaluated by assessing its Type I error rates and power under specified conditions. The standardized latent mean difference effect size measure ($\hat{\delta}_\kappa$), which had been proposed to assess the magnitude of a latent mean difference estimate, was evaluated through examining its parameter bias and relative parameter bias under specified conditions. In addition, the performance of model fit indices, including the χ^2 test of model fit, RMSEA, CFI, TLI and SRMR, in terms of the correct and incorrect model rejection rates were investigated. Conditions that were manipulated in this simulation study include the sample size ratio, latent mean difference magnitude, loading difference magnitude, factor loading pattern and factor variance ratio.

In this chapter, Type I error rates of the LRT_κ are presented first. Next, the power of the LRT_κ is presented. Third, parameter bias of the LRT_κ is presented for conditions when the true latent mean difference was equal to zero, followed by the relative parameter bias of the LRT_κ for conditions when the true latent mean difference was equal to 0.5. Last but not the least, model rejection rates of the χ^2 test of model fit, RMSEA, CFI, TLI and SRMR are reported. For each model fit index, model rejection rates in conditions where the true latent mean difference was

equal to zero or 0.5 are presented in separate tables. In addition, model rejection rates of the RMSEA, CFI, TLI and SRMR using different cutoff values are shown in separate tables.

Type I Error Rates of the LRT_{κ}

In this simulation study, the Type I error rate of the LRT_{κ} is defined as the proportion of the rejections of the null hypothesis of equal latent means across groups, out of the 1,000 converged replications, in conditions where the true latent mean difference is equal to zero. Table 5 presents the observed Type I error rates of the LRT_{κ} under varying conditions. Values above the dashed line are the Type I error rates in the equal factor loading conditions whereas values below the dashed line are the Type I error rates in the unequal factor loading conditions. In each design cell, two Type I error rates are presented. The first one is the Type I error rate observed when implementing the RI strategy. The second one is the Type I error rate obtained when the factor-variance scaling method was used. In this simulation study, Bradley's (1978) liberal criterion of $\alpha \pm 1/2\alpha$ was used to evaluate Type I error rates using a nominal alpha level of 0.05. If Type I error rates were less than 2.5%, then they were considered overly conservative and are marked in boldface in Table 5. If Type I error rates were higher than 7.5%, then they were considered overly liberal and are underlined. Table 6 contains the explanations of abbreviations used in Table 5 and all other tables in this dissertation.

Equal Factor Loading Pattern

In the equal factor loading conditions, all observed Type I error rates when the RI strategy was used were within the criterion of 0.05 ± 0.025. Type I error rates did not vary substantially or systematically as a function of the sample size ratio or the factor variance ratio. When implementing the factor-variance scaling method, one Type I error rate was beyond the

criterion of 0.05 ± 0.025. This overly liberal Type I error rate (0.079) was found in the condition where the sample size ratio was 1:4 and the factor variance ratio was 1.2:0.8.

Table 5

Type I Error Rates of the Likelihood Ratio Test

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	0.056	0.057	0.049	0.051	0.058	0.058
		100:400	0.062	0.052	0.060	0.079	0.059	0.044
		400:100	0.068	0.067	0.051	0.037	0.055	0.070
0.1	1st Loading	250:250	0.047	0.045	0.050	0.049	0.057	0.058
	Unequal	100:400	0.047	0.043	0.046	0.052	0.068	0.046
		400:100	0.043	0.044	0.062	0.054	0.056	0.068
	2nd Loading	250:250	0.046	0.046	0.046	0.046	0.050	0.051
	Unequal	100:400	0.052	0.051	0.047	0.058	0.059	0.047
		400:100	0.038	0.045	0.052	0.040	0.064	0.083
	All Lower	250:250	0.068	0.067	0.057	0.060	0.071	0.071
		100:400	0.056	0.049	0.054	0.064	0.045	0.029
		400:100	0.058	0.070	0.049	0.038	0.051	0.068
	Mixed	250:250	0.048	0.048	0.054	0.055	0.047	0.046
		100:400	0.050	0.049	0.049	0.059	0.052	0.037
		400:100	0.053	0.054	0.056	0.040	0.059	0.081
0.4	1st Loading	250:250	0.044	0.043	0.058	0.059	0.044	0.042
	Unequal	100:400	0.048	0.030	0.053	0.044	0.060	**0.019**
		400:100	0.058	0.070	0.054	0.054	0.061	0.096
	2nd Loading	250:250	0.053	0.050	0.054	0.054	0.055	0.053
	Unequal	100:400	0.055	0.027	0.053	0.050	0.049	**0.025**
		400:100	0.050	0.064	0.045	0.043	0.051	0.085
	All Lower	250:250	0.055	0.050	0.048	0.044	0.050	0.040
		100:400	0.044	**0.016**	0.066	0.046	0.061	**0.016**
		400:100	0.045	0.082	0.047	0.059	0.051	0.113
	Mixed	250:250	0.055	0.056	0.052	0.052	0.042	0.041
		100:400	0.050	0.041	0.045	0.058	0.040	**0.021**
		400:100	0.038	0.029	0.045	**0.024**	0.046	0.052

Note. Type I error rates smaller than 0.025 are shown in boldface. Type I error rates greater than 0.075 are underlined. Abbreviations used in this table are explained in Table 6.

Table 6

Explanations of Abbreviations Used in the Tables in this Dissertation

Abbreviation	Explanation
RI	Reference indicator strategy
FV	Factor-variance-based scaling method
Equal Loading	All factor loadings were generated to be equal across groups
1st Loading Unequal	The first factor loading (RI) was generated to have a higher true value in group two than in group one with the condition-specific loading difference
2nd Loading Unequal	The second factor loading was generated to have a higher true value in group two than in group one with the condition-specific loading difference
All Lower	Both the first (RI) and second factor loading were generated to have higher true values in group two than in group one with the condition-specific loading difference
Mixed	The first factor loading (RI) was generated to have a higher true value in group one and the second factor loading to have a higher true value in group two with the condition-specific loading difference

Unequal Factor Loading Pattern

In the unequal factor loading conditions, the Type I error rates of the LRT_{κ} when the RI strategy was implemented were within the criterion of 0.05 ± 0.025. When the factor-variance scaling method was used, twelve observed Type I error rates were beyond the criterion of 0.05 ± 0.025. There were one overly conservative and one overly liberal Type I error rates under the 1:1 factor variance ratio conditions. One overly conservative Type I error rate was found in the 1.2:0.8 factor variance ratio conditions. Additionally, four overly conservative and five overly liberal Type I error rates were found in conditions in which the factor variance ratio was 0.8:1.2. The Type I error rates that were beyond the criterion all occurred in the unequal sample size conditions (with a sample size ratio of 1:4 or 4:1). In addition, most of the Type I error rates that exceeded the criterion were found in conditions in which the loading difference was equal to 0.4.

Power of the LRT_{κ}

In the present study, the power of the LRT_{κ} is defined as the proportion of the rejections of the null hypothesis of equal latent means across groups, out of the 1,000 converged replications, in conditions where the true latent mean difference is equal to 0.5. Table 7 presents the observed power rates of the LRT_{κ}. A criterion of 0.90 (Goodman & Berlin, 1994) was used to evaluate the power of the LRT_{κ} in this simulation study, meaning that if power rates were lower than 0.90 then they were considered too low and are underlined in Table 7.

Table 7

Power of the Likelihood Ratio Test

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	0.979	0.979	0.983	0.983	0.982	0.982
		100:400	0.906	0.906	0.866	0.892	0.911	0.893
		400:100	0.888	0.891	0.920	0.894	0.873	0.893
0.1	1st Loading	250:250	0.987	0.988	0.983	0.982	0.980	0.981
	Unequal	100:400	0.924	0.926	0.927	0.937	0.941	0.921
		400:100	0.898	0.905	0.929	0.916	0.897	0.927
	2nd Loading	250:250	0.984	0.984	0.988	0.988	0.982	0.982
	Unequal	100:400	0.925	0.917	0.890	0.911	0.920	0.890
		400:100	0.904	0.912	0.947	0.936	0.912	0.937
	All Lower	250:250	0.991	0.991	0.993	0.993	0.994	0.994
		100:400	0.935	0.928	0.939	0.942	0.953	0.926
		400:100	0.919	0.931	0.938	0.927	0.916	0.929
	Mixed	250:250	0.983	0.984	0.985	0.986	0.986	0.986
		100:400	0.906	0.901	0.890	0.903	0.908	0.892
		400:100	0.911	0.908	0.942	0.924	0.892	0.910
0.4	1st Loading	250:250	0.998	0.998	1.000	1.000	0.999	0.999
	Unequal	100:400	0.971	0.967	0.975	0.978	0.989	0.971
		400:100	0.964	0.973	0.973	0.979	0.940	0.962
	2nd Loading	250:250	0.997	0.997	0.999	0.999	0.998	0.998
	Unequal	100:400	0.984	0.967	0.976	0.979	0.984	0.966
		400:100	0.965	0.974	0.974	0.974	0.944	0.962
	All Lower	250:250	1.000	1.000	0.999	0.999	1.000	1.000
		100:400	0.993	0.980	0.993	0.989	0.995	0.981
		400:100	0.982	0.995	0.993	0.995	0.971	0.990
	Mixed	250:250	0.995	0.996	0.995	0.997	0.998	0.998
		100:400	0.961	0.957	0.953	0.964	0.972	0.959
		400:100	0.925	0.920	0.931	0.906	0.924	0.932

Note. Power rates smaller than 0.90 are underlined. Abbreviations used in this table are explained in Table 6.

Equal Factor Loading Pattern

In the equal factor loading conditions, three power rates when the RI strategy was implemented were lower than the criterion of 0.90. In the condition where the factor variance ratio was 1:1 and sample size ratio was 4:1, the power rate was equal to 0.888. In the condition where the factor variance ratio was 1.2:0.8 and the sample size ratio was 1:4, the power rate was equal to 0.866. Another low power rate (0.873) was found in the condition where the factor variance ratio was 0.8:1.2 and the sample size ratio was 4:1. Although these values were lower than the criterion of 0.90, they did not deviate substantially from 0.90. In addition, all these low power rates were found in the unequal sample size conditions with a sample size ratio of 1:4 or 4:1. Power rates were higher in the equal sample size conditions. For the 1:1, 1:4 and 4:1 sample size ratio conditions, average power rates were 0.981, 0.894 and 0.894, respectively. In addition, power rates when the RI strategy was implemented did not vary substantially or systematically as a function of the factor variance ratio. Average power rates were 0.924, 0.923 and 0.922 for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively.

In the equal factor loading conditions, five out of nine power rates when the factor-variance scaling method was used were lower than the criterion of 0.90. However, they did not deviate substantially from 0.90 (ranging from 0.866 to 0.894). All of the low power rates were found in the unequal sample size conditions. Power rates were higher in the equal sample size conditions than in the unequal sample size conditions. Average power rates were 0.981, 0.897 and 0.893 for the 1:1, 1:4, and 4:1 sample size ratio conditions, respectively. Additionally, power rates based on the factor-variance scaling method did not differ substantially or systematically as a function of the factor variance ratios.

Unequal Factor Loading Pattern

In the unequal factor loading conditions, five power rates that were based on the RI strategy were lower than the criterion of 0.90. They all occurred in conditions in which the loading difference was equal to 0.1 and the sample size ratio was 1:4 or 4:1. Although these power rates were below the criterion, they were very close to the criterion (in the range of 0.89 to 0.898). Across the two loading difference magnitudes, power rates based on the RI strategy were slightly higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Average power rates were 0.942 and 0.978, respectively, for the 0.1 and 0.4 loading difference conditions. Across the three sample size ratios, power rates based on the RI strategy were slightly higher in the equal sample size conditions than in the unequal sample size conditions. Three perfect power rates were observed in equal sample size conditions and the rest of the power rates in equal sample size conditions were also close to 1.00. For the 1:1, 1:4 and 4:1 sample size ratio conditions, average power rates were 0.992, 0.95 and 0.937, respectively. In addition, power rates based on the RI strategy did not vary substantially or systematically as a function of the factor variance ratios or factor loading patterns. Average power rates were 0.958, 0.963 and 0.958 for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively, and were 0.960, 0.960, 0.972 and 0.948 for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively.

When the factor-variance scaling method was implemented, two observed power rates were lower than the criterion of 0.90, although they did not differ substantially from the criterion. These two low power rates were found in conditions in which the loading difference was 0.1, the factor variance ratio was 0.8:1.2 and the sample size ratio was 1:4. The trends in the power rates based on the factor-variance scaling method were consistent with those based on the RI strategy.

Specifically, power rates were slightly higher in the 0.4 loading difference conditions (with a mean of 97.8%) than in the 0.1 loading difference conditions (with a mean of 94.2%). Across the three sample size ratio conditions, the equal sample size condition led to slightly higher power rates than did the unequal sample size conditions. Average power rates were 99.3%, 94.4% and 94.3% for the respective 1:1, 1:4 and 4:1 sample size ratio conditions. Additionally, power rates did not show large differences across the three factor variance ratios or the four factor loading patterns.

Parameter Bias of the $\widehat{\delta}_{\kappa}$

In this simulation study, parameter bias of the standardized latent mean difference effect size measure ($\hat{\delta}_{\kappa}$) was calculated in conditions where the true latent mean difference was equal to zero. Table 8 shows the parameter bias of the $\hat{\delta}_{\kappa}$ under varying conditions. A cutoff value of 0.05 was used to evaluate the acceptability of the parameter bias, meaning that parameter bias with the absolute value less than 0.05 indicated acceptable bias.

Inspecting the parameter bias in Table 8, no parameter bias' absolute value was found greater than 0.05, regardless of the factor scaling method used. Parameter bias values ranged from 0.013 to 0.014 with a mean of -0.0003 and a standard deviation of 0.006. Only 10 out of 144 parameter bias' absolute values were greater than 0.01 (but still lower than 0.05). Both negative and positive parameter bias were observed. However, no clear trend was found.

Table 8

Parameter Bias of the Standardized Latent Mean Difference Effect Size Measure

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	-0.003	-0.003	-0.003	-0.003	0.002	0.002
		100:400	-0.005	-0.005	-0.008	-0.008	-0.008	-0.008
		400:100	0.002	0.002	0.003	0.003	-0.001	-0.001
0.1	1st Loading	250:250	0.000	0.0001	0.006	0.006	-0.007	-0.007
	Unequal	100:400	0.014	0.014	0.001	0.001	-0.009	-0.009
		400:100	0.002	0.002	0.004	0.004	0.007	0.007
	2nd Loading	250:250	0.001	0.001	0.001	0.001	-0.010	-0.010
	Unequal	100:400	0.008	0.008	0.001	0.001	0.0001	0.001
		400:100	0.009	0.009	0.003	0.003	0.002	0.002
	All Lower	250:250	-0.001	-0.001	-0.009	-0.009	0.002	0.002
		100:400	-0.005	-0.005	0.011	0.011	0.013	0.013
		400:100	-0.005	-0.005	0.003	0.003	0.004	0.004
	Mixed	250:250	-0.010	-0.010	0.001	0.001	-0.003	-0.003
		100:400	0.002	0.002	-0.013	-0.013	0.003	0.003
		400:100	0.0002	0.0002	-0.002	-0.001	-0.006	-0.006
0.4	1st Loading	250:250	-0.001	-0.001	0.004	0.004	-0.007	-0.007
	Unequal	100:400	0.005	0.005	0.006	0.006	0.011	0.011
		400:100	-0.001	-0.002	-0.003	-0.003	0.002	0.002
	2nd Loading	250:250	-0.005	-0.005	-0.009	-0.008	-0.002	-0.001
	Unequal	100:400	0.005	0.005	0.010	0.010	-0.005	-0.006
		400:100	0.002	0.002	0.005	0.005	0.001	0.001
	All Lower	250:250	0.004	0.004	-0.006	-0.006	0.006	0.006
		100:400	-0.010	-0.010	0.002	0.002	-0.005	-0.005
		400:100	-0.006	-0.006	0.002	0.002	-0.001	-0.002
	Mixed	250:250	-0.008	-0.008	0.002	0.002	0.004	0.004
		100:400	-0.011	-0.011	-0.004	-0.004	-0.003	-0.003
		400:100	-0.005	-0.005	0.003	0.003	-0.002	-0.002

Note. Abbreviations used in this table are explained in Table 6.

Relative Parameter Bias of the $\widehat{\delta}_{\kappa}$

In conditions where the true latent mean difference was equal to 0.5, relative parameter bias of the $\hat{\delta}_{\kappa}$ was calculated. Table 9 presents the relative parameter bias of the $\hat{\delta}_{\kappa}$ for each combination of the factor variance ratio, loading difference magnitude, factor loading pattern and

sample size ratio conditions. When using Hoogland and Boomsma's (1998) cutoff value of 0.05, relative parameter bias values that were greater than 0.05 represented unacceptable relative parameter bias. These values are underlined in Table 9.

Table 9

Relative Parameter Bias of the Standardized Latent Mean Difference Effect Size Measure

			Factor Variance Ratio					
			1:1		1.2:0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	-0.008	-0.008	0.012	0.012	-0.004	-0.003
		100:400	-0.011	-0.011	0.004	0.005	-0.013	-0.012
		400:100	-0.006	-0.006	0.004	0.004	-0.004	-0.008
0.1	1st Loading	250:250	0.017	0.016	0.022	0.024	0.033	0.033
	Unequal	100:400	0.019	0.019	0.050	0.051	0.013	0.014
		400:100	0.050	0.050	0.049	0.051	0.037	0.037
	2nd Loading	250:250	0.013	0.013	0.034	0.035	0.008	0.009
	Unequal	100:400	0.010	0.010	-0.004	-0.003	-0.014	-0.013
		400:100	0.041	0.041	0.050	0.051	0.057	0.057
	All Lower	250:250	0.046	0.045	0.059	0.061	0.037	0.036
		100:400	0.015	0.014	0.042	0.043	0.010	0.010
		400:100	0.066	0.066	0.074	0.076	0.070	0.068
	Mixed	250:250	0.0004	0.0002	-0.013	-0.011	0.015	0.015
		100:400	-0.002	-0.002	-0.017	-0.017	-0.015	-0.014
		400:100	0.009	0.009	0.016	0.017	0.014	0.011
0.4	1st Loading	250:250	0.111	0.103	0.155	0.154	0.097	0.088
	Unequal	100:400	0.039	0.034	0.052	0.051	0.043	0.037
		400:100	0.215	0.204	0.229	0.229	0.175	0.155
	2nd Loading	250:250	0.129	0.121	0.147	0.145	0.092	0.083
	Unequal	100:400	0.025	0.021	0.077	0.076	0.038	0.032
		400:100	0.198	0.187	0.213	0.214	0.180	0.162
	All Lower	250:250	0.184	0.153	0.230	0.212	0.130	0.091
		100:400	0.060	0.046	0.068	0.059	0.033	0.019
		400:100	0.344	0.313	0.390	0.374	0.299	0.255
	Mixed	250:250	-0.005	-0.006	-0.011	0.008	0.030	0.022
		100:400	0.008	0.006	0.050	0.058	-0.007	-0.012
		400:100	-0.064	-0.058	-0.087	-0.072	-0.015	-0.026

Note. Relative parameter bias values equal to or greater than 0.05 are underlined. Abbreviations used in this table are explained in Table 6.

Equal Factor Loading Pattern

In the equal factor loading conditions, all relative parameter bias' absolute values were lower than the cutoff value of 0.05, regardless of the factor scaling method used. In addition, relative parameter bias based on the RI strategy and based on the factor-variance scaling method showed consistent trends: negative relative parameter bias occurred in conditions where the factor variance ratio was 1:1 or 0.8:1.2 and positive relative parameter bias appeared in conditions where the factor variance ratio was 1.2:0.8. Although the relative parameter bias values were in opposite directions, their absolute values did not differ substantially across factor variance ratios or sample size ratios.

Unequal Factor Loading Pattern

When implementing the RI strategy, the relative parameter bias' absolute values exceeded the cutoff value in 33 conditions. More unacceptable relative parameter bias was found in conditions in which the loading difference was 0.4 than in conditions in which the loading difference was 0.1. Relative parameter bias was also more substantial in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Average relative parameter bias values were 0.025 and 0.107 respectively for the 0.1 and 0.4 loading difference conditions. For the four factor loading patterns, more unacceptable relative parameter bias values were found in the "all lower" pattern conditions than in the "1st loading unequal," "2nd loading unequal" and "mixed" pattern conditions. Additionally, relative parameter bias was more substantial in the "all lower" pattern conditions than in the other three pattern conditions. Relative parameter bias values were lowest in the "mixed" pattern conditions. Average relative parameter bias values were 0.078, 0.072, 0.12 and -0.005 for the "1st loading unequal," "2nd loading unequal," "all

lower" and "mixed "pattern conditions, respectively. Across the three sample size ratios, relative parameter bias values were lowest in the 1:4 sample size ratio conditions and were more substantial in 4:1 sample size ratio conditions than in the other two sample size ratio conditions. Average relative parameter bias values were 0.065, 0.025 and 0.109 for the 1:1, 1:4 and 4:1 sample size ratio conditions, respectively. Regarding relative parameter bias across the three factor variance ratios, no clear trend was found.

An analysis of variance (ANOVA) was conducted to investigate each design factor's main and interaction effects. The dependent variable was the relative parameter bias of the $\hat{\delta}_\kappa$, and independent variables included the loading difference magnitude, factor loading pattern, sample size ratio and factor variance ratio. Before conducting the ANOVA, the normality of the distribution of the relative parameter bias was checked. Both skewness and kurtosis values were less in magnitude than 0.4, supporting the assumption of normality for the relative parameter bias of the $\hat{\delta}_\kappa$ estimates. Although there were four independent variables, only two-way interactions were analyzed to facilitate explanation for the results. Given the large sample size being analyzed, practical measures of effect were evaluated for significance. The effect size that was used was the partial η^2. A minimum cutoff value of 0.06 was used to indicate practical significance (Cohen, 1988). Table 10 contains ANOVA test results for relative parameter bias that was based the RI strategy. None of the main or interaction effect sizes were greater than 0.06.

Table 10

ANOVA of the Relative Parameter Bias of the Standardized Latent Mean Difference Effect Size Measure When Using the RI Strategy

Source	Sum of Squares	DF	Mean Square	Partial η^2
Two-Way Interaction				
Loading Difference*Loading Pattern	60.285	3	20.095	0.010
Loading Difference*Sample Size Ratio	28.446	2	14.223	0.005
Loading Difference* Factor Variance Ratio	2.184	2	1.092	0.000
Loading Pattern*Sample Size Ratio	58.869	6	9.811	0.010
Loading Pattern*Factor Variance Ratio	5.974	6	0.996	0.001
Sample Size Ratio* Factor Variance Ratio	0.331	4	0.083	0.000
Main Effect				
Loading Difference	119.829	1	119.829	0.020
Loading Pattern	146.775	3	48.925	0.024
Sample Size Ratio	84.628	2	42.314	0.014
Factor Variance Ratio	5.768	2	2.884	0.001

When the factor variance scaling method was implemented, values of relative parameter bias closely matched those found when using the RI strategy. In addition, the trends found for sources of the relative parameter bias were consistent with those found when using the RI strategy. An ANOVA was also conducted to assess each factor's main and interaction effects using relative parameter bias based on the factor-variance scaling method as the dependent variable. Note that again, the skewness and kurtosis for the distribution of the relative parameter bias were found to support the assumption of normality. Table 11 presents the ANOVA test results. Similar to the effect sizes in Table 10, no main or interaction effect sizes in Table 11 were greater in magnitude than the criterion of 0.06.

Table 11

ANOVA of the Relative Parameter Bias of the Standardized Latent Mean Difference Effect Size Measure When Using the Factor-Variance Scaling Method

Source	Sum of Squares	DF	Mean Square	Partial η^2
Two-Way Interaction				
Loading Difference*Loading Pattern	44.415	3	14.805	0.008
Loading Difference*Sample Size Ratio	24.713	2	12.356	0.004
Loading Difference* Factor Variance Ratio	4.936	2	2.468	0.001
Loading Pattern*Sample Size Ratio	53.642	6	8.940	0.009
Loading Pattern*Factor Variance Ratio	5.475	6	0.913	0.001
Sample Size Ratio* Factor Variance Ratio	0.097	4	0.024	0.000
Main Effect				
Loading Difference	95.365	1	95.365	0.016
Loading Pattern	120.185	3	40.062	0.020
Sample Size Ratio	77.688	2	38.844	0.013
Factor Variance Ratio	11.143	2	5.571	0.002

Model Rejection Rates Associated with the χ^2 Test of Model Fit

In this simulation study, the performance of the χ^2 test of model fit was evaluated by assessing its model rejection rates under varying conditions. Table 12 contains the model rejection rates of the χ^2 test in conditions where the true latent mean difference was equal to zero. Table 13 presents the model rejection rates in conditions where the true latent mean difference was equal to 0.5. In each table, the values above the dashed line are the model rejection rates for conditions with data generated to have equal factor loadings across groups. When using the RI strategy, the estimating models were correctly specified in the equal factor loading conditions. Thus, the incorrect model rejection rates can be interpreted as Type I error rates of the χ^2 test. When the factor-variance scaling method was implemented, model rejection rates under the equal factor variance conditions can also be interpreted as Type I error rates; model rejection

rates under the unequal factor variance conditions can be interpreted as statistical power, since the estimating models were incorrectly specified by constraining unequal factor variances to a value of one across groups. In each table, the values below the dashed line are the model rejection rates for conditions with data generated to have unequal factor loading patterns although the estimating models assumed the loadings to be equal across groups. Thus, the model rejection rates under the dashed line represent the statistical power of the χ^2 test. In Tables 12 and 13, all incorrect model rejection rates that can be interpreted as Type I error rates are italicized.

Latent Mean Difference of Zero

Equal factor loading pattern. Table 12 contains the model rejection rates of the χ^2 test in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, model rejection rates when the RI strategy was used were all within Bradley's (1978) criterion of 0.05 ± 0.025, and did not differ systematically as a function of the factor variance ratios or sample size ratios. In the equal factor variance conditions, model rejection rates based on the factor-variance scaling method were also within the criterion of 0.05 ± 0.025. In conditions in which the factor variance ratio was 1.2:0.8 or 0.8:1.2, model rejection rates based on the factor-variance scaling method, which represented the power of the χ^2 test, were found to be very low (ranging from 8% to 12%).

Table 12

Model Rejection Rates of the χ^2 Test of Model Fit in Conditions Where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.062*	*0.067*	*0.039*	0.094	*0.060*	0.120
		100:400	*0.062*	*0.057*	*0.055*	0.092	*0.050*	0.080
		400:100	*0.063*	*0.069*	*0.060*	0.087	*0.060*	0.093
0.1	1st Loading Unequal	250:250	0.062	0.064	0.071	0.108	0.049	0.167
		100:400	0.066	0.073	0.068	0.082	0.054	0.102
		400:100	0.061	0.059	0.055	0.076	0.076	0.126
	2nd Loading Unequal	250:250	0.058	0.060	0.054	0.083	0.055	0.155
		100:400	0.071	0.075	0.054	0.071	0.069	0.136
		400:100	0.052	0.055	0.071	0.090	0.063	0.134
	All Lower	250:250	0.089	0.091	0.066	0.086	0.079	0.233
		100:400	0.075	0.087	0.054	0.066	0.073	0.169
		400:100	0.079	0.089	0.072	0.089	0.073	0.164
	Mixed	250:250	0.079	0.085	0.075	0.136	0.078	0.149
		100:400	0.090	0.085	0.064	0.094	0.094	0.141
		400:100	0.094	0.070	0.069	0.116	0.068	0.110
0.4	1st Loading Unequal	250:250	0.244	0.400	0.249	0.242	0.223	0.761
		100:400	0.156	0.258	0.166	0.165	0.119	0.448
		400:100	0.217	0.302	0.175	0.167	0.201	0.514
	2nd Loading Unequal	250:250	0.240	0.431	0.247	0.245	0.221	0.739
		100:400	0.140	0.244.	0.184	0.181	0.123	0.457
		400:100	0.192	0.277	0.170	0.178	0.198	0.544
	All Lower	250:250	0.378	0.902	0.485	0.613	0.274	0.988
		100:400	0.190	0.546	0.219	0.307	0.158	0.821
		400:100	0.407	0.803	0.454	0.518	0.384	0.955
	Mixed	250:250	0.863	0.864	0.811	0.895	0.808	0.897
		100:400	0.616	0.616	0.688	0.714	0.445	0.649
		400:100	0.601	0.619	0.466	0.636	0.674	0.696

Note. Model rejection rates that are equal to Type I error rates are italicized. The rest of the model rejection rates can be interpreted as statistical power. Abbreviations used in this table are explained in Table 6.

Unequal factor loading pattern. In the unequal factor loading conditions, correct model rejection rates when the RI strategy was implemented varied as a function of the loading difference magnitude. In conditions in which the true loading difference was 0.1, all model

rejection rates based on the RI strategy were below 10%. In conditions in which the true loading difference was 0.4, model rejection rates were between 11.9% and 86.3%. Average rejection rates were 6.9% and 33.8% for the 0.1 and 0.4 loading difference conditions, respectively. Across the three factor variance ratios, no substantial differences were observed. Average rejection rates were 21.3%, 21.2% and 19.4% respectively for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Regarding the model rejection rates across the four factor loading patterns, different trends were observed for the 0.1 and 0.4 loading difference conditions. Within conditions with a loading difference of 0.1, model rejection rates in the four loading pattern conditions did not differ substantially or systematically. Within conditions with a loading difference of 0.4, model rejection rates were generally higher in the "mixed" pattern conditions than in the other loading pattern conditions. The model rejection rates in the "mixed" pattern conditions ranged from 44.5% to 86.3% with a mean of 66.4%. The model rejection rates in the "all lower" pattern conditions (with a mean of 32.8%) were lower than those in the "mixed" pattern conditions but were higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. The model rejections rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with means of 19.4% and 19.1%, respectively). For the model rejection rates across the three sample size ratios, clear trends were only observed in conditions in which the loading difference was 0.4. To be precise, model rejection rates were generally higher in the equal sample size conditions (with a mean of 42.0%) than in the unequal sample size conditions. In addition, model rejection rates were generally higher in the unequal 4:1 sample size ratio conditions (with a mean of 34.5%) than in the unequal 1:4 sample size ratio conditions (with a mean of 26.7%).

Table 12 also contains the model rejection rates when implementing the factor-variance scaling method. In the unequal factor loading conditions, model rejection rates based on the factor-variance scaling method were higher in conditions with a loading difference of 0.4 than in conditions with a loading difference of 0.1. Several high model rejection rates (e.g., 90.2%, 98.8% and 95.5%) were found only in the 0.4 loading difference conditions. The average rejection rate was 10.9% for the 0.1 loading difference conditions and increased to 58.1% as the magnitude of the loading difference increased to 0.4. For the three factor variance ratios, the factor variance ratio of 0.8:1.2 generally led to higher model rejection rates than did the factor variance ratios of 1:1 and 1.2:0.8. Within conditions with a loading difference of 0.1, model rejection rates under the 0.8:1.2 factor variance ratio conditions were all above 10% whereas under the other two ratio conditions, most of the model rejection rates were below 10%. Within conditions with a loading difference of 0.4, all model rejection rates in the 0.8:1.2 factor variance ratio conditions were above 40% and two of them exceeded 95% (i.e., 98.8% and 95.5%). For the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, average model rejection rates were 30%, 24.8% and 42.7%, respectively. Regarding the model rejection rates across the four factor loading patterns, different trends were observed in the 0.1 and 0.4 loading difference conditions. In conditions in which the true loading difference was 0.1, model rejection rates did not differ substantially or systematically across the four loading patterns. In conditions in which the true loading difference was 0.4, model rejection rates were generally higher in the "all lower" and "mixed" pattern conditions (with means of 71.7% and 73.2%, respectively) than in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with means of 36.2% and 36.6%, respectively). Within conditions with a loading difference of 0.4, most of the rejection rates in the "all lower" and "mixed" pattern conditions were above 60% whereas most of the rejection rates in the "1st

loading unequal" and "2nd loading unequal" conditions were below 40%. In addition, model rejection rates did not differ much in the "1st loading unequal" and "2nd loading unequal" pattern conditions.

Model rejection rates based on the factor-variance scaling method were also compared across the three sample size ratios. Similar trends were observed to those found when implementing the RI strategy. Specifically, within conditions with a loading difference of 0.1, model rejection rates did not differ greatly in the three sample size ratio conditions. Within conditions with a loading difference of 0.4, model rejection rates were generally higher in the equal sample size conditions than in the two unequal sample size conditions. Additionally, model rejection rates were generally higher in the 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. For the 1:1, 1:4 and 4:1 sample size ratio conditions, average model rejection rates were 66.5%, 45.1% and 51.7%, respectively.

Latent Mean Difference of 0.5

Table 13 presents the model rejection rates of the χ^2 test in conditions where the true latent mean difference was equal to 0.5. Compared to the values in Table 12, it was found that model rejection rates in conditions in which the true latent mean difference was equal to 0.5 were slightly higher than those in conditions in which the true latent mean difference was equal to zero.

Equal factor loading pattern. In the equal factor loading conditions, incorrect model rejection rates when the RI strategy was implemented, which can be interpreted as Type I error rates of the χ^2 test, were all within the criterion of 0.05 ± 0.025, and did not differ systematically as a function of the factor variance ratios or sample size ratios. In the equal factor variance conditions, incorrect model rejection rates based on the factor-variance scaling method

were also within the criterion of 0.05 ± 0.025. In the unequal factor variance conditions, correct model rejection rates based on the factor-variance scaling method, which can be interpreted as statistical power of the χ^2 test, were found to be low (ranging from 8.5% to 10.9%).

Unequal factor loading pattern. In conditions in which the true latent mean difference was equal to 0.5, trends in the model rejection rates across the loading difference magnitudes, factor variance ratios and factor loading patterns were consistent with those in conditions in which the true latent mean difference was equal to zero. First, when using the RI strategy, model rejection rates in the 0.4 loading difference conditions (with a mean of 38.3%) were higher than those in the 0.1 loading difference conditions (with a mean of 7%). Model rejection rates in the 0.4 loading difference conditions were in the range of 11.0% to 91.4% whereas model rejection rates in the 0.1 loading difference conditions were in the range of 5.2% to 9.0%. High model rejection rates (e.g., 99.1% and 97.3%) were found only in the 0.4 loading difference conditions. Second, model rejection rates did not differ substantially as a function of the factor variance ratio. Average rejection rates were 23%, 24.4% and 20.6%, respectively, for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2. Third, the trends across the four factor loading patterns were different for the 0.1 and 0.4 loading difference conditions. In conditions in which the true loading difference was 0.1, model rejection rates did not vary substantially or systematically across the four factor loading patterns. In conditions in which the true loading difference was 0.4, model rejection rates were generally higher in the "mixed" pattern conditions (with a mean of 66.4%) than in the other three loading pattern conditions. The "all lower" pattern led to lower rejection rates (with a mean of 32.8%) than did the "mixed" pattern but produced slightly higher rejection rates than did the "1st loading unequal" and "2nd loading unequal" patterns (with means of 19.4% and 19.1%, respectively).

Table 13

Model Rejection Rates of the χ^2 Test of Model Fit in Conditions Where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.059*	*0.056*	*0.055*	0.109	*0.054*	0.106
		100:400	*0.059*	*0.066*	*0.057*	0.085	*0.056*	0.090
		400:100	*0.069*	*0.066*	*0.062*	0.104	*0.050*	0.092
0.1	1st Loading Unequal	250:250	0.063	0.074	0.065	0.107	0.065	0.161
		100:400	0.071	0.078	0.060	0.083	0.080	0.146
		400:100	0.067	0.069	0.064	0.091	0.060	0.123
	2nd Loading Unequal	250:250	0.071	0.073	0.081	0.115	0.073	0.165
		100:400	0.054	0.054	0.063	0.088	0.070	0.121
		400:100	0.069	0.067	0.070	0.093	0.062	0.118
	All Lower	250:250	0.075	0.085	0.079	0.097	0.068	0.230
		100:400	0.066	0.083	0.052	0.076	0.067	0.179
		400:100	0.061	0.072	0.073	0.090	0.071	0.157
	Mixed	250:250	0.078	0.082	0.080	0.159	0.076	0.167
		100:400	0.075	0.085	0.075	0.108	0.074	0.118
		400:100	0.090	0.087	0.065	0.107	0.086	0.128
0.4	1st Loading Unequal	250:250	0.236	0.401	0.262	0.272	0.232	0.757
		100:400	0.131	0.233	0.153	0.154	0.120	0.457
		400:100	0.270	0.359	0.274	0.280	0.238	0.594
	2nd Loading Unequal	250:250	0.270	0.447	0.307	0.307	0.184	0.758
		100:400	0.142	0.238	0.185	0.178	0.110	0.462
		400:100	0.256	0.344	0.232	0.236	0.246	0.599
	All Lower	250:250	0.401	0.875	0.477	0.608	0.289	0.991
		100:400	0.188	0.552	0.220	0.299	0.124	0.813
		400:100	0.492	0.815	0.548	0.614	0.381	0.973
	Mixed	250:250	0.914	0.908	0.904	0.950	0.866	0.944
		100:400	0.622	0.641	0.722	0.737	0.476	0.671
		400:100	0.754	0.761	0.733	0.842	0.822	0.842

Note. Model rejection rates that are equal to Type I error rates are italicized. The rest of the model rejection rates are equal to statistical power. Abbreviations used in this table are explained in Table 6.

All of these trends (across loading difference magnitudes, factor variance ratios and factor loading patterns) were consistent with those observed in Table 12 in which the true latent mean difference was equal to zero. However, in conditions in which the true latent mean

difference was equal to 0.5, the trends across the three sample size ratios were slightly different from those in conditions in which the true latent mean difference was equal to zero. More specifically, within conditions with a loading difference of 0.4, model rejection rates were consistently higher in the 1:1 and 4:1 sample size ratio conditions (with means of 44.5% and 43.7%, respectively) than in the 1:4 conditions (with a mean of 26.6%). Additionally, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. This trend was different from that observed in conditions with a latent mean difference of zero in which model rejection rates were generally higher in the 1:1 sample size ratio conditions than in the 4:1 and 1:4 sample size ratio conditions.

In conditions where the true latent mean difference was equal to 0.5, model rejection rates based on the factor-variance scaling method are also reported in Table 13. Across the two magnitudes of the loading difference, model rejection rates were higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.1, most of the model rejection rates were around 10% and the highest rejection rate was 23%. Within conditions with a loading difference of 0.4, most of the model rejection rates exceeded 50% and several of them were close to 100%. For the loading difference conditions of 0.1 and 0.4, average rejection rates were 10.9% and 58.1%, respectively. Across the three factor variance ratios, model rejection rates were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. Within conditions with a loading difference of 0.1, all model rejection rates under the 0.8:1.2 conditions were above 11%. In contrast, all model rejection rates under the equal factor variance conditions were below 9% and half of the model rejection rates under the 1.2:0.8 factor variance ratio conditions were below 10%. Within conditions with a loading difference of 0.4, all model

rejection rates under the 0.8:1.2 factor variance ratio conditions were above 45% and five of them exceeded 80%. On the other hand, most of the model rejection rates under the 1:1 and 1.2:0.8 factor variance ratio conditions were around 20%. For the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, average rejection rates were 31.2%, 27.9% and 44.5%, respectively. Regarding the model rejection rates across the four factor loading patterns, the trends were also different in the 0.1 and 0.4 loading difference conditions. When the true loading difference was 0.1, model rejection rates were similar in the four loading pattern conditions with means of 10.4%, 9.9%, 11.9% and 11.6%, respectively. When the true loading difference was 0.4, model rejection rates were generally higher in the "all lower" and "mixed" pattern conditions(with means of 72.7% and 81.1%, respectively) than in the "1st loading unequal" and "2nd loading unequal" pattern conditions(with means of 39% and 39.7%, respectively). Across the three sample size ratios, no clear trend was observed in conditions with a loading difference of 0.1. However, in conditions with a loading difference of 0.4, model rejection rates varied as a function of the sample size ratios. Specifically, the sample size ratios of 1:1 and 4:1 always yielded higher rejection rates than did the sample size ratio of 1:4. In addition, the equal sample size conditions generally led to higher rejection rates than did the sample size ratio conditions of 4:1. Within conditions with a loading difference of 0.4, average rejection rates were 68.5%, 60.5% and 45.3% respectively for the 1:1, 4:1 and 1:4 sample size ratio conditions.

Model Rejection Rates of the RMSEA

In the present study, the RMSEA model fit index was evaluated by assessing its model rejection rates under varying conditions. Two RMSEA cutoff values, 0.05 and 0.06, which were suggested by Steiger (1989) and Hu and Bentler (1999), respectively, were used to determine whether the null hypothesis of model fit should be rejected. If the RMSEA value was greater

than the relevant cutoff (0.05 or 0.06) then the null hypothesis of model fit was rejected. Table 14 and Table 15 present the model rejection rates of the RMSEA using cutoff values of 0.05 and 0.06, respectively, in conditions where the true latent mean difference was equal to zero. Table 16 and Table 17 contain the model rejection rates of the RMSEA using cutoff values of 0.05 and 0.06, respectively, in conditions where the true latent mean difference was equal to 0.5. In each table, values above the dashed line are the model rejection rates in the equal factor loading conditions. When using the RI strategy, incorrect model rejection rates of the RMSEA can be interpreted similarly to Type I error rates since the estimating models were correctly specified. When using the factor-variance scaling method, model rejection rates under the equal factor variance conditions can also be interpreted similarly to Type I error rates. On the other hand, model rejection rates in the unequal factor variance conditions (with a factor variance ratio of 1.2:0.8 or 0.8:1.2), can be interpreted similarly to statistical power because the estimating models were incorrectly specified by constraining unequal factor variances to a value of one across groups. In each table, values below the dashed line are the model rejection rates of the RMSEA under the unequal factor loading conditions. These rejection rates can be interpreted similarly to statistical power since the estimating models were incorrectly specified by constraining all factor loadings to be equal across groups. In Tables 14 to 17, all incorrect rejection rates that can be interpreted similarly to Type I error rates are italicized.

Latent Mean Difference of Zero

In conditions in which the true latent mean difference was equal to zero, model rejection rates observed when using cutoff values of 0.05 and 0.06 were compared across loading difference magnitudes, factor variance ratios, factor loading patterns, and sample size ratios.

Table 14

Model Rejection Rates of the RMSEA When Using a Cutoff of 0.05 in Conditions Where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.023*	*0.020*	*0.015*	0.041	*0.018*	0.045
		100:400	*0.027*	*0.022*	*0.024*	0.039	*0.023*	0.033
		400:100	*0.027*	*0.027*	*0.025*	0.034	*0.027*	0.053
0.1	1st Loading Unequal	250:250	0.035	0.029	0.031	0.050	0.025	0.083
		100:400	0.031	0.032	0.023	0.028	0.023	0.054
		400:100	0.024	0.026	0.022	0.031	0.035	0.079
	2nd Loading Unequal	250:250	0.026	0.032	0.021	0.039	0.025	0.081
		100:400	0.031	0.030	0.020	0.027	0.028	0.062
		400:100	0.027	0.025	0.035	0.049	0.027	0.063
	All Lower	250:250	0.037	0.039	0.031	0.044	0.041	0.143
		100:400	0.043	0.042	0.022	0.033	0.035	0.089
		400:100	0.039	0.037	0.033	0.038	0.025	0.097
	Mixed	250:250	0.036	0.033	0.036	0.073	0.037	0.068
		100:400	0.038	0.034	0.028	0.050	0.039	0.073
		400:100	0.038	0.037	0.038	0.059	0.021	0.043
0.4	1st Loading Unequal	250:250	0.151	0.268	0.156	0.146	0.125	0.616
		100:400	0.088	0.148	0.094	0.087	0.066	0.287
		400:100	0.118	0.181	0.100	0.094	0.116	0.363
	2nd Loading Unequal	250:250	0.157	0.273	0.156	0.157	0.122	0.625
		100:400	0.077	0.159	0.111	0.113	0.063	0.307
		400:100	0.100	0.148	0.102	0.098	0.109	0.393
	All Lower	250:250	0.258	0.808	0.341	0.478	0.172	0.976
		100:400	0.098	0.378	0.130	0.185	0.081	0.689
		400:100	0.264	0.671	0.320	0.377	0.231	0.918
	Mixed	250:250	0.774	0.770	0.704	0.814	0.705	0.808
		100:400	0.475	0.481	0.578	0.595	0.327	0.500
		400:100	0.460	0.470	0.331	0.493	0.542	0.558

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Using a cutoff value of 0.05 in the equal factor loading conditions. Table 14 presents the model rejection rates of the RMSEA when using a cutoff of 0.05 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, five out of nine

incorrect model rejection rates based on the RI strategy were overly conservative(e.g.,1.5%, 1.8% and 2.3%). When the factor-variance scaling method was used, two incorrect model rejection rates in the equal factor variance conditions were overly conservative (2.0% and 2.2%). Model rejection rates in the unequal factor variance conditions, which equated to statistical power, were found to be low (in the range of 3.3% to 5.3%).

Using a cutoff value of 0.05 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were lower in the 0.1 loading difference conditions than in the 0.4 loading difference conditions. Model rejection rates within conditions with a loading difference of 0.1 were in the range of 2.1% to 4.3%. In contrast, within conditions with a loading difference of 0.4, almost all the model rejection rates were above 10% and several of them were greater than 45% with the highest rejection rate of 77.4%. The average rejection rate was 3.1% for the 0.1 loading difference conditions and increased to 24.5% in the 0.4 loading difference conditions. Across the three factor variance ratios, model rejection rates did not vary substantially or systematically. Average rejection rates were 14.3%, 14.4% and 12.6% for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively. Regarding the model rejection rates across the four factor loading patterns, no clear trend was observed in the 0.1 loading difference conditions. Model rejection rates, however, varied as a function of the factor loading patterns in the 0.4 loading difference conditions. Specifically, model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with means of 11.3% and 11.1%, respectively), and increased slightly in the "all lower" pattern conditions (with a mean of 21.1%). Highest model rejection rates were found in the "mixed" pattern conditions in which more than half of the rejection rates were above 50%. The average rejection rate was 54.4% for the "mixed" pattern conditions. Regarding the trends

across the three sample size ratios, clear trends were also found in the 0.4 loading difference conditions. Specifically, the sample size ratio conditions of 1:1 consistently led to higher model rejection rates than did the sample size ratio conditions of 1:4. The sample size ratio conditions of 1:1 also generally led to higher model rejection rates than did the sample size ratio conditions of 4:1. In addition, in most of the conditions, the sample size ratio conditions of 4:1 led to higher model rejection rates than did the sample size ratio conditions of 1:4. For the 1:1, 4:1 and 1:4 sample size ratio conditions, average rejection rates were 31.8%, 23.3% and 18.2%, respectively.

Model rejection rates based on the factor-variance scaling method for the unequal factor loading conditions are also reported in Table 14. Consistent with the trends observed when using the RI strategy, model rejection rates based on the factor-variance scaling method were higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.1, almost all the model rejection rates were below 10%. In contrast, within conditions with a loading difference of 0.4, most of the model rejection rates were greater than 30%. A few high rejection rates (e.g., 91.8% and 97.6%) were found only in the 0.4 loading difference conditions. Average rejection rates were 5.1% and 42.9%, respectively, for the 0.1 and 0.4 loading difference conditions. Across the three factor variance ratios, model rejection rates based on the factor-variance scaling method were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance conditions. This trend was most obvious within conditions with a loading difference of 0.4, where all the model rejection rates under the 0.8:1.2 facto variance ratio conditions were above 30% and more than half of them exceeded 80%. The two highest rejection rates (97.6% and 91.8%) were also found under the 0.8:1.2 factor variance ratio conditions. Regarding the model rejection rates across the four factor loading patterns, clear trends were found only in conditions in which the loading

difference was 0.4. Within conditions with a loading difference of 0.1, model rejection rates did not vary substantially or systematically as a function of the factor loading patterns. Within conditions with a loading difference of 0.4, model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions with average rejection rates of 24.3% and 25.3%, respectively. Higher rejection rates were observed in the "all lower" and "mixed" pattern conditions with average rejection rates of 60.9% and 61%, respectively. The trends across the three sample size ratios were also different in the 0.1 and 0.4 loading difference conditions. Substantial differences across the three sample size ratios were only found in the 0.4 loading difference conditions. Specifically, model rejection rates were consistently higher in the equal sample size conditions than in the unequal 1:4 and 4:1 sample size ratio conditions. In addition, model rejection rates were generally higher in the 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. Average rejection rates were 56.2%, 39.7% and 32.7% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Using a cutoff value of 0.06 in the equal factor loading conditions. Table 15 presents the model rejection rates of the RMSEA when using a cutoff of 0.06 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, all model rejection rates obtained when using a cutoff value of 0.06 were lower than their counterparts in Table 14 in which a cutoff value of 0.05 was used. When the RI strategy was implemented, the incorrect model rejection rates of the RMSEA, which can be interpreted similarly to Type I error rates, were found to be overly conservative (ranging from 0.1% to 0.5%). When the factor-variance scaling method was used, the incorrect model rejection rates under the equal factor variance conditions were also overly conservative (ranging from 0.1% to 0.5%). The model

rejection rates under the unequal factor variance conditions, which can be interpreted similarly to statistical power, were found to be very low (in the range of 0.6% to 1%).

Table 15

Model Rejection Rates of the RMSEA When Using a Cutoff of 0.06 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.004*	*0.005*	*0.004*	0.010	*0.004*	0.010
		100:400	*0.001*	*0.001*	*0.003*	0.007	*0.003*	0.007
		400:100	*0.005*	*0.004*	*0.002*	0.006	*0.004*	0.010
0.1	1st Loading Unequal	250:250	0.000	0.000	0.003	0.007	0.003	0.015
		100:400	0.003	0.003	0.000	0.005	0.003	0.009
		400:100	0.002	0.003	0.003	0.004	0.007	0.021
	2nd Loading Unequal	250:250	0.003	0.002	0.005	0.010	0.004	0.014
		100:400	0.007	0.006	0.004	0.002	0.005	0.011
		400:100	0.006	0.006	0.003	0.004	0.004	0.011
	All Lower	250:250	0.005	0.007	0.002	0.007	0.004	0.047
		100:400	0.007	0.005	0.006	0.007	0.005	0.022
		400:100	0.004	0.006	0.008	0.007	0.004	0.018
	Mixed	250:250	0.008	0.010	0.003	0.018	0.006	0.011
		100:400	0.009	0.008	0.006	0.014	0.005	0.012
		400:100	0.009	0.010	0.005	0.012	0.006	0.010
0.4	1st Loading Unequal	250:250	0.039	0.091	0.037	0.035	0.040	0.362
		100:400	0.019	0.035	0.020	0.019	0.010	0.116
		400:100	0.035	0.051	0.031	0.025	0.034	0.174
	2nd Loading Unequal	250:250	0.042	0.109	0.047	0.040	0.037	0.374
		100:400	0.022	0.041	0.038	0.034	0.015	0.105
		400:100	0.032	0.047	0.024	0.020	0.034	0.166
	All Lower	250:250	0.097	0.561	0.144	0.234	0.050	0.910
		100:400	0.022	0.145	0.031	0.057	0.017	0.402
		400:100	0.103	0.417	0.123	0.165	0.093	0.757
	Mixed	250:250	0.549	0.532	0.473	0.605	0.458	0.601
		100:400	0.249	0.247	0.299	0.322	0.148	0.245
		400:100	0.217	0.220	0.147	0.253	0.276	0.295

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Using a cutoff value of 0.06 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates observed when using a cutoff value of 0.06 were lower than those in Table 14 in which a cutoff value of 0.05 was used. When implementing the RI strategy, the trends in the model rejection rates when using a cutoff of 0.06 were consistent with those observed when using a cutoff value of 0.05. First, model rejection rates were consistently higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.1, two model rejection rates based on the RI strategy were equal to zero and the rest of them were all below 1%. The average rejection rate in the 0.1 loading difference conditions was 0.5%. Within conditions with a loading difference of 0.4, model rejection rates were in the range of 1.7% to 54.9% with an average rejection rate of 11.3%. Second, model rejection rates did not differ substantially or systematically as a function of the factor variance ratios. For the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, average rejection rates were 6.2%, 6.1% and 5.3%, respectively. Third, clear trends across the four factor loading patterns were only found in conditions with a loading difference of 0.4. More specifically, all model rejection rates in the "mixed" pattern conditions were higher than those in the"1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. Model rejection rates in the "mixed" pattern conditions were all above 14% with an average rejection rate of 31.3%. Model rejection rates in the "all lower" pattern conditions were lower than those in the "mixed" pattern conditions but were slightly higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Additionally, model rejection rates were similar in the latter two pattern conditions. Average rejection rates were 7.6%, 2.9% and 3.2% for the "all lower," "1st loading unequal" and "2nd loading unequal" pattern conditions, respectively. Finally, clear trends across the three sample size ratios were also only found in

conditions in which the loading difference was 0.4. More specifically, model rejection rates under the 1:1 sample size ratio conditions (with an average rejection rate of 16.8%) were generally higher than those under the 4:1 and 1:4 sample size ratio conditions (with average rejection rates of 9.6% and 7.4%, respectively) . Additionally, model rejection rates were generally higher in the 4:1 sample size ratios conditions than in the 1:4 conditions.

When the factor-variance scaling method was implemented, model rejection rates using a cutoff value of 0.06 showed similar trends as those using a cutoff value of 0.05. First, model rejection rates based on the factor-variance scaling method increased as the loading difference magnitude increased. In conditions in which the loading difference was 0.1, one rejection rate was equal to zero and the rest of them were in the range of 0.2% to 4.7% with an average rejection rate of 1%. In conditions in which the loading difference was 0.4, model rejection rates were in the range of 1.9% to 91% and many of them exceeded 10%. The average rejection rate was 24.5% for the 0.4 loading difference conditions. Second, model rejection rates based on the factor-variance scaling method were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. Average rates were 10.7%, 7.9% and 19.6%, respectively, for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Third, the trends in the model rejection rates across the four factor loadings were different in the 0.1 and 0.4 loading difference conditions. In conditions with a loading difference of 0.1, model rejection rates were all low and did not vary substantially or systematically as a function of the factor loading patterns. In conditions with a loading difference of 0.4, model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with average rejection rates of 10.1% and 10.4%, respectively). However, higher model rejection rates were found in the "all lower" and "mixed" pattern conditions with average rejection rates of 40.5%

and 36.9%, respectively. Last but not the least, only in conditions in which the true loading difference was 0.4, clear trends across the three sample size ratios were observed. More specifically, model rejection rates in the equal 1:1 sample size ratio conditions were higher than those in the unequal 1:4 and 4:1 sample size ratio conditions. Additionally, in most of the conditions, the 4:1 sample size ratio conditions led to higher rejection rates than did the 1:4 sample size ratio conditions.

Latent Mean Difference of 0.5

When the true latent mean difference was equal to 0.5, model rejection rates of the RMSEA were investigated when varying the loading difference magnitude, factor variance ratio, factor loading pattern and sample size ratio. In addition, model rejection rates of the RMSEA were compared when using cutoff values of 0.05 and 0.06, respectively.

Using a cutoff value of 0.05 in the equal factor loading conditions. Table 16 presents the model rejection rates of the RMSEA when using a cutoff of 0.05 in condition in which the true latent mean difference was equal to 0.5. In the equal factor loading conditions, five out of nine incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were beyond the criterion of 0.05 ± 0.025 and were found to be overly conservative. All of these overly conservative rejection rates occurred in the unequal factor variance conditions. When the factor-variance scaling method was used, the incorrect rejection rates in the equal factor variance conditions were within the criterion. However, the correct model rejection rates in the unequal factor variance conditions, which can be interpreted similarly to statistical power of the RMSEA to correctly reject fit of the incorrect model, were found to be very low (in the range of 3.5% to 4.8%).

Table 16

Model Rejection Rates of the RMSEA When Using a Cutoff of 0.05 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.028*	*0.026*	*0.017*	0.048	*0.020*	0.046
		100:400	*0.026*	*0.026*	*0.020*	0.039	*0.024*	0.042
		400:100	*0.031*	*0.026*	*0.026*	0.041	*0.024*	0.035
0.1	1st Loading Unequal	250:250	0.028	0.029	0.035	0.051	0.028	0.090
		100:400	0.023	0.025	0.021	0.027	0.044	0.075
		400:100	0.031	0.034	0.031	0.043	0.026	0.061
	2nd Loading Unequal	250:250	0.031	0.034	0.032	0.054	0.033	0.092
		100:400	0.025	0.021	0.025	0.039	0.032	0.060
		400:100	0.031	0.028	0.028	0.038	0.018	0.053
	All Lower	250:250	0.036	0.040	0.038	0.046	0.032	0.136
		100:400	0.030	0.028	0.021	0.028	0.027	0.087
		400:100	0.028	0.027	0.038	0.044	0.033	0.087
	Mixed	250:250	0.033	0.034	0.041	0.077	0.036	0.084
		100:400	0.031	0.030	0.024	0.047	0.043	0.073
		400:100	0.045	0.044	0.031	0.047	0.040	0.058
0.4	1st Loading Unequal	250:250	0.136	0.265	0.163	0.150	0.144	0.634
		100:400	0.071	0.132	0.087	0.085	0.064	0.303
		400:100	0.157	0.226	0.168	0.152	0.141	0.449
	2nd Loading Unequal	250:250	0.166	0.308	0.185	0.176	0.102	0.637
		100:400	0.074	0.137	0.095	0.090	0.057	0.301
		400:100	0.144	0.215	0.143	0.142	0.149	0.445
	All Lower	250:250	0.275	0.778	0.340	0.456	0.184	0.975
		100:400	0.106	0.384	0.133	0.186	0.062	0.685
		400:100	0.354	0.719	0.391	0.459	0.276	0.935
	Mixed	250:250	0.850	0.842	0.830	0.900	0.751	0.870
		100:400	0.466	0.484	0.575	0.582	0.350	0.525
		400:100	0.638	0.639	0.584	0.727	0.700	0.729

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Using a cutoff value of 0.05 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were lower in conditions

in which the true loading difference was 0.1 than in conditions in which the true loading difference was 0.4. In the 0.1 loading difference conditions, all model rejection rates based on the RI strategy were lower than 5%. In the 0.4 loading difference conditions, model rejection rates increased. Most of the model rejection rates exceeded 10% and several of them were above 80%. Average model rejection rates were 3.1% and 28.1% for the respective 0.1 and 0.4 loading difference conditions. In the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, model rejection rates based on the RI strategy were similar with average rejection rates of 15.9%, 16.9% and 14.1%, respectively. Inspecting the model rejection rates across the four factor loading patterns, different trends were observed in the 0.1 and 0.4 loading difference conditions. In conditions in which the loading difference was 0.1, model rejection rates in the four loading pattern conditions did not differ substantially or systematically. In conditions in which the loading difference was 0.4, model rejection rates were higher in the "mixed" pattern conditions than in the other three pattern conditions. Most of the rejection rates in the "mixed" pattern conditions were above 50% with the average rejection rate of 63.8%. Although model rejection rates in the "all lower" pattern conditions were lower than those in the "mixed" pattern conditions, they were slightly higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. In addition, model rejection rates did were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions. For the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, average rejection rates were 12.6%, 12.4%, 23.6% and 63.8%, respectively. The trends across the three sample size ratios were also different in the 0.1 and 0.4 loading difference conditions. When the true loading difference was set to a value of 0.1, there was no clear trend across the three sample size ratios. When the true loading difference was set to a value of 0.4, the sample size ratio conditions of 1:1

and 4:1 led to higher rejection rates than did the 1:4 conditions. Model rejection rates in the 1:1 and 4:1 sample size ratio conditions were all greater than 15% whereas several low rejection rates (e.g., 5.7% and 6.2%) were found in the 1:4 sample size ratio conditions. Within conditions with a loading difference of 0.4, average rejection rates were 34.4%, 32% and 17.8% respectively for the 1:1, 4:1 and 1:4 sample size ratio conditions.

When using a cutoff value of 0.05, model rejection rates based on the factor-variance scaling method are also reported in Table 16. In the unequal factor loading conditions, model rejection rates based on the factor-variance scaling method increased while the loading difference magnitude increased. In conditions in which the loading difference was 0.1, all model rejection rates based on the factor-variance scaling method were below 10%. In conditions in which the loading difference was 0.4, almost all the model rejection rates were above 10% and three of them were equal to or greater than 90%. For the 0.1 and 0.4 loading difference conditions, average rejection rates were 5.2% and 46.5%, respectively. Across the three factor variance ratios, model rejection rates based on the factor-variance scaling method were generally higher in conditions in which the factor variance ratio was 0.8:1.2 (with a mean of 35.2%) than in conditions in which the factor variance ratio was 1:1 or 1.2:0.8 (with means of 22.9% and 19.4%, respectively). Two highest rejection rates (97.5% and 93.5%) were found under the 0.8:1.2 factor variance ratio conditions. In addition, model rejection rates were similar in the 1:1 and 1.2:0.8 factor variance ratio conditions. Regarding the model rejection rates across the four factor loading patterns, clear trends were only observed in conditions with a loading difference of 0.4. To be precise, model rejection rates were higher in the "all lower" and "mixed" pattern conditions than in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Within conditions with a loading difference of 0.4, average rejection rates were 70%, 62%, 26.6% and

27.2% for the respective "mixed," "all lower," "1st loading unequal" and "2nd loading unequal" pattern conditions. Across the three sample size ratios, model rejection rates based on the factor-variance scaling method were generally higher in the equal 1:1 sample size ratio conditions than in the unequal 1:4 and 4:1 sample size ratio conditions. For the two unequal sample size conditions, model rejection rates were generally higher in conditions with a sample size ratio of 4:1 than in conditions with a sample size ratio of 1:4. These trends across the sample size ratios were most obvious in conditions with a loading difference of 0.4 in which average rejection rates were 58.3%, 48.6% and 32.5% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Using a cutoff value of 0.06 in the equal factor loading conditions. Table 17 presents the model rejection rates of the RMSEA when using a cutoff of 0.06 in condition where the true latent mean difference was equal to 0.5. In the equal factor loading conditions, model rejection rates observed when using a cutoff value of 0.06 were lower than their counterparts in Table 16 in which a cutoff value of 0.05 was used. When the RI strategy was implemented, the incorrect model rejection rates of the RMSEA, which can be interpreted similarly to Type I error rates, all exceeded 0.05 ± 0.025 and were found to be overly conservative. When the factor-variance scaling method was used, incorrect model rejection rates under the equal 1:1 factor variance ratio conditions, which can also be interpreted similarly to Type I error rates, were found to be overly conservative. In the unequal 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, the correct model rejection rates (or the power rates) of the RMSEA were in the range of 0.5% to 0.7%, indicating extremely low power.

Table 17

Model Rejection Rates of the RMSEA When Using a Cutoff of 0.06 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.004*	*0.001*	*0.005*	0.006	*0.003*	0.007
		100:400	*0.005*	*0.005*	*0.005*	0.007	*0.003*	0.005
		400:100	*0.004*	*0.001*	*0.003*	0.009	*0.006*	0.006
0.1	1st Loading Unequal	250:250	0.005	0.005	0.004	0.010	0.004	0.023
		100:400	0.007	0.003	0.002	0.008	0.006	0.013
		400:100	0.006	0.005	0.007	0.006	0.006	0.015
	2nd Loading Unequal	250:250	0.002	0.003	0.008	0.014	0.003	0.021
		100:400	0.002	0.002	0.003	0.007	0.005	0.014
		400:100	0.004	0.003	0.005	0.002	0.002	0.011
	All Lower	250:250	0.006	0.007	0.002	0.006	0.007	0.045
		100:400	0.004	0.005	0.003	0.002	0.007	0.019
		400:100	0.003	0.003	0.008	0.009	0.006	0.024
	Mixed	250:250	0.010	0.005	0.009	0.018	0.006	0.022
		100:400	0.009	0.010	0.003	0.008	0.008	0.013
		400:100	0.006	0.007	0.008	0.012	0.009	0.012
0.4	1st Loading Unequal	250:250	0.034	0.091	0.044	0.040	0.044	0.385
		100:400	0.017	0.040	0.023	0.021	0.019	0.118
		400:100	0.033	0.065	0.041	0.038	0.031	0.216
	2nd Loading Unequal	250:250	0.052	0.112	0.063	0.059	0.026	0.356
		100:400	0.015	0.031	0.025	0.025	0.008	0.108
		400:100	0.037	0.065	0.048	0.040	0.038	0.207
	All Lower	250:250	0.110	0.535	0.158	0.242	0.058	0.912
		100:400	0.020	0.155	0.042	0.056	0.010	0.403
		400:100	0.139	0.480	0.159	0.212	0.111	0.807
	Mixed	250:250	0.654	0.641	0.643	0.727	0.508	0.649
		100:400	0.254	0.248	0.323	0.341	0.143	0.258
		400:100	0.401	0.399	0.318	0.459	0.427	0.470

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Using a cutoff value of 0.06 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates when using a cutoff value of 0.06 were lower than those in Table 16 in which a cutoff value of 0.05 was used. In addition, when using a cutoff

value of 0.06, the trends in the model rejection rates across loading difference magnitudes, factor loading patterns and factor variance ratios were consistent with those found when using a cutoff value of 0.05. First, model rejection rates based on the RI strategy were found to be higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.1, most of the model rejection rates were lower than 1%. Within conditions with a loading difference of 0.4, higher rejection rates were observed, particularly in the "mixed" pattern conditions in which model rejection rates were in the range of 14.3% to 65.4%. For the loading difference conditions of 0.1 and 0.4, average rejection rates were 0.5% and 14.1%, respectively. Second, model rejection rates when the RI strategy was used did not vary substantially or systematically across the three factor variance ratios. Average rejection rates were 7.6%, 8.1% and 6.2% for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively. Third, the trends across the four factor loading patterns were different when the true loading difference was set to 0.1 or 0.4. In conditions in which the true loading difference was 0.1, model rejection rates were similar in the four loading pattern conditions. Model rejection rates, however, varied as a function of the factor loading patterns in conditions in which the true loading difference was 0.4. Specifically, the "mixed" pattern conditions, in which all the model rejection rates were above 14%, led to the highest rejection rates. The model rejection rates in the "all lower" pattern conditions were lower than those in the "mixed" pattern conditions but were higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions, although the difference was not large. Within conditions with a loading difference of 0.4, average rejection rates were 40.8% , 9%, 3.5% and 3.2% for the "mixed," "all lower," "1st loading unequal" and "2nd loading unequal" pattern conditions, respectively. When using a cutoff value of 0.06, the trends across the three sample size ratios were slightly different from those

observed when using a cutoff value of 0.05. Within conditions with a loading difference of 0.4, the sample size ratio of 1:1 consistently led to higher rejection rates than did the sample size ratio of 1:4. In addition, under most of the conditions, the sample size ratio of 1:1 led to higher rejection rates than did the sample size ratio of 4:1. When using a cutoff value of 0.05, the model rejection rates in the sample size ratio conditions of 1:1 and 4:1 did not show substantial or systematic differences.

When using a cutoff value of 0.06, model rejection rates based on the factor-variance scaling method were also lower than those obtained when using a cutoff value of 0.05. In addition, the trends in the model rejection rates were consistent with those found when using a cutoff of 0.05. As expected, model rejection rates based on the factor-variance scaling method were higher in conditions with a loading difference of 0.4 than in conditions with a loading difference of 0.1. Average rejection rates were 1.1% and 27.8% for the 0.1 and 0.4 loading difference conditions, respectively. Across the three factor variance ratios, model rejection rates based on the factor-variance scaling method were generally higher under the 0.8:1.2 factor variance ratio conditions than under the 1:1 and 1.2:0.8 factor variance ratio conditions. Two highest rejection rates (i.e., 91.2% and 80.7%) were found in the 0.8:1.2 factor variance ratio conditions. Average rejection rates were 12.2%, 9.8% and 21.3% respectively for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Regarding the rejection rates across the four factor loading patterns, clear trends were only found in conditions in which the true loading difference was 0.4. Specifically, model rejection rates in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with means of 11.3% and 11.1%, respectively) were lower than those in the "all lower" and "mixed" pattern conditions (with means of 42.2% and 46.6%, respectively). Additionally, model rejection rates were similar in the first two loading pattern conditions and

did not differ substantially in the latter two conditions. Within conditions with a loading difference of 0.4, model rejection rates also differed as a function of the sample size ratios. The equal sample size conditions consistently led to higher rejection rates than did the unequal sample size conditions. For the two unequal sample size conditions, the sample size ratio of 4:1 consistently produced higher model rejection rates than did the sample size ratio of 1:4. Within the 0.4 loading difference conditions, average rejection rates were 39.6%, 28.8% and 15% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Model Rejection Rates of the CFI

In this simulation study, the performance of the CFI model fit index in terms of its model rejection rates was investigated under a variety of conditions. Two CFI cutoff values, 0.90 and 0.95, which were proposed by Bentler and Bonnet (1980) and Hu and Bentler (1999), respectively, were used to determine whether the null hypothesis of model fit should be rejected. If the CFI value was less than the relevant cutoff value (0.90 or 0.95) then the null hypothesis of model fit was rejected. Table 18 and Table 19 contain the model rejection rates of the CFI when using cutoff values of 0.90 and 0.95, respectively, in conditions where the true latent mean difference was equal to zero. Table 20 and Table 21 present the model rejection rates of the CFI when using cutoff values of the 0.90 and 0.95, respectively, in conditions where the true latent mean difference was equal to 0.5. In each table, values above the dashed line are the model rejection rates in the equal factor loading conditions. When the RI strategy was implemented, the incorrect model rejection rates of the CFI can be interpreted similarly to Type I error rates. When the factor-variance scaling method was used, the model rejection rates in the equal factor variance conditions can also be interpreted similarly to Type I error rates. The model rejection rates under the unequal factor variance conditions, on the other hand, can be interpreted similarly

to statistical power. Values below the dashed line are the model rejection rates of the CFI in the unequal factor loading conditions. Since the estimating models were incorrectly specified by constraining unequal factor loadings to be equal across groups, model rejection rates below the dashed lines can be interpreted similarly to statistical power. In Tables 18 to 21, all incorrect rejection rates that can be interpreted similarly to Type I error rates are italicized.

Latent Mean Difference of Zero

Using a cutoff value of 0.90 in the equal factor loading conditions. Table 18 presents the model rejection rates of the CFI using a cutoff value of 0.90 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, three out of nine incorrect model rejection rates that were based on the RI strategy differed substantially from 5%. Two of them were found to be overly high (7.5% and 8%) and the third one was overly low (2.1%). All of them occurred in the unequal sample size conditions. In the equal factor loading conditions, the incorrect rejection rates based on the factor-variance scaling method did not differ substantially from 5%. All correct rejection rates based on the factor-variance scaling method, were found to be overly low (in the range of 3.6% to 13%).

Using a cutoff value of 0.90 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were generally higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. However, there were a few exceptions. In conditions in which the sample size ratio was 1:4 and factor loadings were in the "1st loading unequal," "2nd loading unequal" or "all lower" patterns, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading

difference conditions. Average rejection rates were 3.8% and 10.6% for the 0.1 and 0.4 loading difference conditions, respectively.

Table 18

Model Rejection Rates of the CFI When Using a Cutoff of 0.90 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.047*	*0.054*	*0.033*	0.083	*0.039*	0.088
		100:400	*0.040*	*0.038*	*0.075*	0.123	*0.026*	0.036
		400:100	*0.042*	*0.049*	*0.021*	0.040	*0.080*	0.130
0.1	1st Loading Unequal	250:250	0.040	0.044	0.039	0.059	0.029	0.107
		100:400	0.034	0.040	0.061	0.078	0.011	0.037
		400:100	0.036	0.043	0.018	0.036	0.083	0.142
	2nd Loading Unequal	250:250	0.039	0.042	0.038	0.051	0.030	0.096
		100:400	0.034	0.037	0.051	0.064	0.016	0.045
		400:100	0.036	0.046	0.026	0.041	0.068	0.149
	All Lower	250:250	0.043	0.050	0.031	0.051	0.023	0.129
		100:400	0.024	0.034	0.035	0.049	0.008	0.038
		400:100	0.060	0.064	0.030	0.037	0.069	0.167
	Mixed	250:250	0.042	0.040	0.037	0.089	0.045	0.082
		100:400	0.045	0.043	0.054	0.081	0.018	0.047
		400:100	0.035	0.034	0.017	0.035	0.054	0.103
0.4	1st Loading Unequal	250:250	0.072	0.160	0.098	0.098	0.042	0.456
		100:400	0.010	0.029	0.038	0.035	0.004	0.044
		400:100	0.127	0.191	0.067	0.066	0.137	0.455
	2nd Loading Unequal	250:250	0.072	0.167	0.087	0.095	0.049	0.450
		100:400	0.019	0.041	0.051	0.055	0.001	0.057
		400:100	0.098	0.161	0.071	0.069	0.147	0.482
	All Lower	250:250	0.026	0.322	0.067	0.119	0.003	0.681
		100:400	0.000	0.005	0.000	0.001	0.000	0.005
		400:100	0.155	0.535	0.162	0.225	0.143	0.886
	Mixed	250:250	0.422	0.437	0.368	0.520	0.346	0.516
		100:400	0.152	0.158	0.301	0.354	0.037	0.097
		400:100	0.139	0.154	0.039	0.100	0.282	0.334

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Across the three factor variance ratios, model rejection rates based on the RI strategy did not differ substantially or systematically as a function of the factor variance ratios. Average rejection rates were 7.3%, 7.4% and 6.9% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Regarding the model rejection rates across the four factor loading patterns, different trends were observed in the 0.1 and 0.4 loading difference conditions. Within conditions with a loading difference of 0.1, model rejection rates in the four loading pattern conditions were similar. Within conditions with a loading difference of 0.4, model rejection rates were higher in the "mixed" pattern conditions (with a mean of 23.2%) than in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions (with means of 6.6%, 6.6% and 6.2%, respectively). In addition, model rejection rates were similar in the latter three loading pattern conditions. The trends in the rejection rates across the three sample size ratios were also different in the 0.1 and 0.4 loading difference conditions. In the 0.1 loading difference conditions, model rejection rates under the three sample size ratios did not show substantial or systematic differences. In the 0.4 loading difference conditions, the 1:1 and 4:1 sample size ratio conditions generally led to higher model rejection rates than did the 1:4 sample size ratio conditions. Comparing the model rejection rates across the 1:1 and 4:1 sample size ratio conditions, no substantial difference was observed. Within conditions with a loading difference of 0.4, average model rejection rates were 13.7%, 13.1% and 5.1% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

In Table 18, model rejection rates based on the factor-variance scaling method when using a cutoff of 0.90 in conditions in which the latent mean difference was equal to 0.5 are presented. Consistent with the trend in the model rejection rates based on the RI strategy, model rejection rates based on the factor-variance scaling method increased as the load difference

magnitude increased with a few exceptions. For example, with the sample size ratio of 1:4 and in the "all lower" pattern conditions, model rejection rates in the 0.4 loading difference conditions were lower than their counterparts in the 0.1 loading difference conditions. For the loading difference conditions of 0.1 and 0.4, average rejection rates were 6.5% and 23.8%, respectively. Across the three factor variance ratios, model rejection rates when the factor-variance scaling method was used were generally higher in the factor variance ratio conditions of 0.8:1.2 than in the factor variance ratio conditions of 1:1 and 1.2:0.8. This trend was most obvious in conditions with a loading difference of 0.4. In conditions in which the true loading difference was 0.4 and the factor variance ratio was 0.8:1.2, most of the model rejection rates were above 40% and the highest rejection rate was 88.6%. In conditions where the true loading difference was 0.4 and the factor variance ratio was 1:1, most of the model rejection rates were below 20% and the highest rate was 53.5%. In conditions where the true loading difference was 0.4 and the factor variance ratio was 1.2:0.8, most of the model rejection rates were below 10% and the highest rate was 52%. Average model rejection rates were 12%, 10% and 23.4%, respectively, for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Across the four factor loading patterns, clear trends were only found in conditions in which the loading difference was 0.4. When the true loading difference was 0.4, model rejection rates were generally higher in the "all lower" and "mixed" pattern conditions than in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Average rejection rates were 17%, 17.5%, 30.9% and 29.7% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. Regarding the model rejection rates across the three sample size ratios, clear trends were also observed only in the 0.4 loading difference conditions. To be precise, model rejection rates were higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 conditions. Additionally, model rejection rates were

similar in the 1:1 and 4:1 sample size ratio conditions. In conditions in which the true loading difference was 0.4, average rejection rate were 33.5%, 30.5% and 7.3% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Using a cutoff value of 0.95 in the equal factor loading conditions. Table 19 presents the model rejection rates of the CFI when using a cutoff value of 0.95 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, the model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were found to be high (ranging from 13.7% to 23.3%). When the factor-variance scaling method was implemented, the incorrect model rejection rates in the equal factor variance conditions, which can also be interpreted similarly to Type I error rates, were high (in the range of 18% to 20.7%). The correct model rejection rates in the unequal factor variance conditions, which can be interpreted similarly to statistical power, were found to below (ranging from 19.5% to 29.9%).

Using a cutoff value of 0.95 in the unequal factor loading conditions. When using a cutoff value of 0.95, model rejection rates in the unequal factor loading conditions were higher than their counterparts in Table 18 in which a cutoff value of 0.90 was used. Additionally, the trends across loading difference magnitudes, factor loading patterns, sample size ratios and factor variance ratios were consistent with those observed when using a cutoff of 0.90. First, model rejection rates based on the RI strategy were generally higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions with a few exceptions. For example, in conditions where the sample size ratio was 1:4 and factor loadings were in the "all lower" pattern, model rejection rates were lower in conditions with a loading difference of 0.4 than in conditions with a loading difference of 0.1. These exceptions were also found in Table 18 when using a

cutoff value of 0.90. For the 0.1 and 0.4 loading difference conditions, average rejection rates were 17.5% and 36.9%, respectively.

Table 19

Model Rejection Rates of the CFI When Using a Cutoff of 0.95 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.200*	*0.207*	*0.164*	0.289	*0.176*	0.278
		100:400	*0.182*	*0.186*	*0.233*	0.299	*0.153*	0.203
		400:100	*0.179*	*0.180*	*0.137*	0.195	*0.220*	0.288
0.1	1st Loading Unequal	250:250	0.173	0.179	0.185	0.257	0.164	0.342
		100:400	0.156	0.170	0.216	0.265	0.126	0.215
		400:100	0.189	0.190	0.151	0.196	0.234	0.352
	2nd Loading Unequal	250:250	0.166	0.163	0.168	0.226	0.166	0.341
		100:400	0.173	0.190	0.214	0.248	0.143	0.230
		400:100	0.192	0.197	0.136	0.194	0.243	0.377
	All Lower	250:250	0.191	0.210	0.166	0.211	0.163	0.395
		100:400	0.153	0.177	0.171	0.196	0.105	0.234
		400:100	0.189	0.205	0.161	0.193	0.216	0.397
	Mixed	250:250	0.190	0.199	0.174	0.316	0.185	0.303
		100:400	0.185	0.187	0.204	0.271	0.148	0.220
		400:100	0.165	0.180	0.131	0.200	0.208	0.296
0.4	1st Loading Unequal	250:250	0.311	0.511	0.357	0.355	0.271	0.826
		100:400	0.163	0.274	0.250	0.251	0.085	0.380
		400:100	0.386	0.478	0.303	0.303	0.393	0.768
	2nd Loading Unequal	250:250	0.314	0.516	0.372	0.381	0.259	0.821
		100:400	0.151	0.279	0.253	0.261	0.084	0.395
		400:100	0.370	0.468	0.280	0.289	0.421	0.779
	All Lower	250:250	0.275	0.841	0.406	0.575	0.140	0.979
		100:400	0.035	0.207	0.098	0.163	0.010	0.374
		400:100	0.518	0.863	0.525	0.605	0.493	0.981
	Mixed	250:250	0.848	0.851	0.803	0.892	0.796	0.889
		100:400	0.568	0.596	0.733	0.768	0.338	0.554
		400:100	0.583	0.598	0.360	0.542	0.729	0.756

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Second, model rejection rates based on the RI strategy did not differ substantially or systematically as a function of the factor variance ratios. Average model rejection rates were 27.7%, 28.4% and 25.5% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Third, when using a cutoff value of 0.95, the trends across the four factor loading patterns were also consistent with those when using a cutoff value of 0.90. Specifically, within conditions with a loading difference of 0.1, model rejection rates were similar in the four loading pattern conditions. Within conditions with a loading difference of 0.4, model rejection rates differed systematically across the four loading patterns. The "mixed" pattern condition generally led to higher model rejection rates than did the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. In the "mixed" pattern conditions, all model rejection rates were above 30% and most of them exceeded 50% with the highest rate of 84.8%. Much lower model rejection rates were found in the"1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 28%, 27.8%, 27.8% and 64% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. Last, the trends across the three sample size ratios were also different in the 0.1 and 0.4 loading difference conditions. When the true loading difference was set to a value of 0.1, no clear trend across the sample size ratios was found. When the true loading difference was set to a value of 0.4, the 1:1 and 4:1 sample size ratio conditions generally led to higher model rejection rates than did the 1:4 sample size ratio conditions. In addition, model rejection rates did not differ substantially or systematically in the 1:1 and 4:1 sample size ratio conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 42.9%, 44.7% and 23.1% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

When using a cutoff value of 0.95, model rejection rates based on the factor-variance scaling method were also higher than those when using a cutoff value of 0.90. The trends across loading difference magnitudes, factor variance ratios, factor loading patterns and sample size ratios were consistent with those found when using a cutoff value of 0.90. For the two loading difference magnitudes, model rejection rates based on the factor-variance scaling method were higher when the loading difference was larger. In the 0.4 loading difference conditions, several model rejection rates were above 80% and two of them were greater than 95%. In contrast, in the 0.1 loading difference conditions, model rejection rates were in the range 16.3% to 39.7%. Average rejection rates were 24.2% and 56.6% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratios, model rejection rates based on the factor-variance scaling method were higher in conditions with a factor variance ratio of 0.8:1.2 than in conditions with a factor variance ratio of 1:1 or 1.2:0.8. All the model rejection rates in the 0.8:1.2 factor variance ratio conditions were above 20% and two of them were greater than 95%. For the 0.8:1.2, 1:1 and 1.2:0.8 factor variance ratio conditions, average model rejection rates were 50.9%, 36.4% and 34.6%, respectively. Regarding the model rejection rates across the four factor loading patterns, clear trends were only found in the 0.4 loading difference conditions. More specifically, model rejection rates in the "all lower" and "mixed" pattern conditions were generally higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Model rejection rates did not differ substantially in the first two pattern conditions and were also similar in the latter two pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 46.1%, 46.5%, 62.1% and 71.6% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. Across the three sample size ratios, the sample size ratio conditions of 1:1 and 4:1

generally led to higher model rejection rates than did the 1:4 sample size ratio conditions. In addition, model rejection rates in the sample size ratio conditions of 1:1 and 4:1 did not show large or systematic differences. These trends were most obvious in the 0.4 loading difference conditions, in which average model rejection rates were 70.3%, 61.9% and 37.5% for the respective 1:1, 4:1 and 1:4 sample size ratio conditions.

Latent Mean Difference of 0.5

Using a cutoff value of 0.90 in the equal factor loading conditions. Table 20 presents the model rejection rates of the CFI when using a cutoff value of 0.90 in conditions where the true latent mean difference was equal to 0.5. In the equal factor loading conditions, the incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were all around 5% with two exceptions (1.9% and 2.2%). These two low rejection rates were found in the unequal sample size conditions. When the factor-variance scaling method was implemented, the incorrect model rejection rates under the 1:1 factor variance ratio conditions were all around 5%. The correct model rejection rates under the 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, which can be interpreted similarly to statistical power, were found to be low (ranging from 4.3% to 10.5%).

Using a cutoff value of 0.90 in the unequal factor loading conditions. In the unequal factor loading conditions, most of the model rejection rates based on the RI strategy were higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. However, there were a few exceptions. For example, in conditions in which the loading difference was 0.4 and the sample size ratio was 1:4, some unexpected low rejection rates (including two rejection rates equal to zero) were observed under the "1st loading unequal," "2nd loading unequal" and

"all lower" pattern conditions. These rejection rates were lower than their counterparts in conditions in which the loading difference was 0.1. Average model rejection rates were 3.9% and 13.9% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratios, model rejection rates based on the RI strategy did not show large differences. Average model rejection rates were 9%, 9.4% and 8.4%, respectively, for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Across the four factor loading patterns, model rejection rates based on the RI strategy did not show clear trends within the 0.1 loading difference conditions. Systematic differences across the four factor loading patterns were found within the 0.4 loading difference conditions. More specifically, model rejection rates in the "mixed" pattern conditions were higher than those in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. Many model rejection rates in the "mixed" pattern condition were above 30% whereas most of the model rejection rates were below 10% in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 8.2%, 7.9%, 7.3% and 32.3% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. Across the three sample size ratios, model rejection rates based on the RI strategy were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. Additionally, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. Average rejection rates were 10.4%, 12% and 4.4% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Table 20

Model Rejection Rates of the CFI When Using a Cutoff of 0.90 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.036*	*0.040*	*0.041*	0.095	*0.031*	0.075
		100:400	*0.046*	*0.046*	*0.062*	0.104	*0.022*	0.048
		400:100	*0.054*	*0.056*	*0.019*	0.043	*0.066*	0.105
0.1	1st Loading Unequal	250:250	0.042	0.050	0.036	0.076	0.033	0.109
		100:400	0.033	0.040	0.052	0.079	0.023	0.043
		400:100	0.048	0.055	0.030	0.047	0.073	0.143
	2nd Loading Unequal	250:250	0.046	0.053	0.042	0.072	0.039	0.114
		100:400	0.030	0.031	0.052	0.082	0.020	0.043
		400:100	0.044	0.047	0.026	0.041	0.064	0.140
	All Lower	250:250	0.039	0.048	0.040	0.055	0.033	0.129
		100:400	0.021	0.027	0.038	0.049	0.011	0.029
		400:100	0.034	0.038	0.029	0.041	0.070	0.166
	Mixed	250:250	0.036	0.036	0.042	0.081	0.042	0.088
		100:400	0.035	0.043	0.061	0.103	0.022	0.043
		400:100	0.045	0.046	0.014	0.032	0.075	0.110
0.4	1st Loading Unequal	250:250	0.069	0.168	0.101	0.103	0.058	0.466
		100:400	0.015	0.042	0.038	0.045	0.004	0.055
		400:100	0.167	0.245	0.114	0.120	0.171	0.536
	2nd Loading Unequal	250:250	0.080	0.184	0.117	0.123	0.033	0.456
		100:400	0.010	0.024	0.051	0.055	0.001	0.053
		400:100	0.143	0.231	0.094	0.110	0.180	0.545
	All Lower	250:250	0.022	0.314	0.063	0.135	0.010	0.693
		100:400	0.000	0.006	0.002	0.005	0.000	0.007
		400:100	0.195	0.594	0.202	0.263	0.166	0.913
	Mixed	250:250	0.547	0.553	0.544	0.669	0.388	0.568
		100:400	0.161	0.181	0.327	0.371	0.049	0.109
		400:100	0.305	0.313	0.135	0.247	0.450	0.512

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

.

Table 20 also contains the model rejection rates based on the factor-variance scaling method when using a cutoff value of 0.90 in conditions where the true latent mean difference was equal to 0.5. Across the two magnitudes of the loading difference, model rejection rates increased as the magnitude of the loading difference increased with a few exceptions. For example, in conditions where the sample size ratio was 1:4 and factor loadings were in the "1st loading unequal," "2nd loading unequal" or "all lower" patterns, model rejection rates decreased as the loading difference magnitude increased. Average model rejection rates were 6.7% and 27.8% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratios, model rejection rates were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. This trend was most obvious in conditions with a loading difference of 0.4. In conditions in which the factor variance ratio was 0.8:1.2 and the loading difference was 0.4, most of the model rejection rates were above 40% and one of them exceeded 90%. In contrast, in conditions in which the true loading difference was 0.4 and the factor variance ratio was 1:1 or 1.2:0.8, most of the model rejection rates were below 30%. Average model rejection rates were 14%, 12.5% and 25.3% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Regarding the model rejection rates across the four factor loading patterns, clear trends were only observed in conditions in which the loading difference was 0.4. Specifically, model rejection rates in the "all lower" and "mixed" pattern conditions were higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Additionally, no systematic difference was found when comparing the model rejection rates in the "all lower" and "mixed" pattern conditions. In conditions in which the true loading difference was 0.4, average model rejection rates were

19.8%, 19.8%, 32.6% and 39.1% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. Across the three sample size ratios, model rejection rates were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. In addition, model rejection rates in the first two sample size ratio conditions did not differ substantially or systematically. Average model rejection rates were 22.3%, 23.1% and 6.5% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Using a cutoff value of 0.95 in the equal factor loading conditions. Table 21 contains the model rejection rates of the CFI when using a cutoff value of 0.95 in conditions where the true latent mean difference was equal to 0.5. In the equal factor loading conditions, model rejection rates obtained when using a cutoff of 0.95 were higher than those in Table 20 in which a cutoff of 0.90 was used. The incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were much higher than 5% (in the range of 13.9% to 21.4%). In the equal factor variance conditions, the incorrect model rejection rates based on the factor-variance scaling method, which can also be interpreted similarly to Type I error rates, were found to be higher than 5% (ranging from 18.1% to 20.4%). In the unequal factor variance conditions (with a factor variance ratio of 1.2:0.8 or 0.8:1.2), the model rejection rates based on the factor-variance conditions can be interpreted similarly to statistical power. These correct model rejection rates were found to be overly low (ranging from 21.2% to 30.8%).

Using a cutoff value of 0.95 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates observed when using a cutoff value of 0.95 were higher than those observed when using a cutoff value of 0.90. In addition, the trends in the model rejection rates across loading difference magnitudes, factor variance ratios, and factor loading patterns were consistent with those obtained when using a cutoff value of 0.90.

When the RI strategy was implemented, model rejection rates were generally higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions, with average rejection rates of 40.8% and 18%, respectively. A few exceptions were observed. For example, in conditions in which the sample size ratio was 1:4, factor variance ratio was 1:1 or 0.8:1.2 and factor loadings were in the "1st loading unequal," "2nd loading unequal" or "all lower" pattern conditions, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. These unexpected results were also observed in Table 20 where a cutoff value of 0.90 was used. Additionally, model rejection rates based on the RI strategy did not differ substantially or systematically as a function of the factor variance ratios. Average model rejection rates were 29.7%, 31.7% and 26.9% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Regarding the model rejection rates across the four factor loading patterns, no clear trend was observed in conditions with a loading difference of 0.1. However, in conditions with a loading difference of 0.4, model rejection rates varied as a function of the factor loading patterns. More specifically, model rejection rates were highest in the "mixed pattern" conditions in which all but one model rejection rates were above 50% and three of them exceeded 80%. Lower and similar rejection rates were found in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions in which most of the rejection rates were below 50% and some of them were lower than 10%. Within conditions with a loading difference of 0.4, average model rejection rates were 31%, 30.7%, 28.6 and 72.9% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively.

Table 21

Model Rejection Rates of the CFI When Using a Cutoff of 0.95 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.176*	*0.181*	*0.181*	0.299	*0.172*	0.285
		100:400	*0.182*	*0.181*	*0.208*	0.283	*0.152*	0.212
		400:100	*0.210*	*0.204*	*0.139*	0.226	*0.214*	0.308
0.1	1st Loading Unequal	250:250	0.180	0.194	0.185	0.257	0.166	0.327
		100:400	0.187	0.198	0.222	0.255	0.150	0.245
		400:100	0.201	0.206	0.154	0.200	0.221	0.342
	2nd Loading Unequal	250:250	0.183	0.199	0.199	0.263	0.166	0.347
		100:400	0.146	0.155	0.206	0.260	0.123	0.213
		400:100	0.188	0.200	0.152	0.205	0.220	0.353
	All Lower	250:250	0.177	0.203	0.192	0.232	0.158	0.399
		100:400	0.152	0.160	0.208	0.237	0.103	0.231
		400:100	0.174	0.198	0.156	0.175	0.229	0.415
	Mixed	250:250	0.195	0.202	0.186	0.314	0.188	0.304
		100:400	0.174	0.177	0.219	0.291	0.138	0.213
		400:100	0.207	0.219	0.141	0.206	0.246	0.326
0.4	1st Loading Unequal	250:250	0.328	0.520	0.413	0.421	0.284	0.815
		100:400	0.139	0.248	0.247	0.256	0.082	0.374
		400:100	0.442	0.546	0.403	0.397	0.452	0.810
	2nd Loading Unequal	250:250	0.356	0.547	0.420	0.428	0.242	0.817
		100:400	0.145	0.259	0.253	0.252	0.071	0.391
		400:100	0.431	0.539	0.357	0.365	0.488	0.815
	All Lower	250:250	0.285	0.815	0.408	0.566	0.152	0.975
		100:400	0.028	0.216	0.013	0.156	0.006	0.375
		400:100	0.578	0.893	0.600	0.675	0.500	0.993
	Mixed	250:250	0.909	0.906	0.897	0.948	0.842	0.945
		100:400	0.583	0.618	0.751	0.771	0.368	0.568
		400:100	0.735	0.743	0.627	0.761	0.849	0.875

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Using a cutoff value of 0.95, the trends across the three sample size ratios were slightly different from those observed when using a cutoff value 0.90. Clear trends across the three sample size ratios were only observed within conditions with a loading difference of 0.4, instead

of in both the 0.1 and 0.4 lading difference conditions. Model rejection rates under the sample size ratio conditions of 1:1 and 4:1 were generally higher than those under the sample size ratio conditions of 1:4. In addition, model rejection rates in the 1:1 and 4:1 sample size ratio conditions did not show systematic differences. Within the 0.4 loading difference conditions, average model rejection rates were 46.1%, 53.9% and 22.4% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

When the factor-variance scaling method was used to set the scale of the latent variable, model rejection rates obtained when using a cutoff value of 0.95 were also higher than those observed when using a cutoff value of 0.90. In addition, the trends across loading difference magnitudes and factor variance ratios were consistent with those using a cutoff value of 0.90. First, as expected, model rejection rates increased as the loading difference magnitude increased. Within the 0.4 loading difference conditions, several model rejection rates were between 80% and 90% or above 90%. In contrast, within the 0.1 loading difference conditions, most of the model rejection rates were between 15% and 35%. Average model rejection rates were 24.8% and 60% for the loading difference conditions of 0.1 and 0.4, respectively. Second, model rejection rates were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. This trend was most obvious in the 0.4 loading difference conditions in which eight out of twelve model rejection rates under the factor variance ratio conditions of 0.8:1.2 were above 80% and three of them were close to 100%.

When using a cutoff value of 0.95, the trends across the four factor loading patterns were not completely consistent with those observed when using a cutoff value of 0.90. Specifically, within conditions with a loading difference of 0.4, model rejection rates were higher in the “all lower” and “mixed” pattern conditions (with average rejection rates of 62.9% and 79.3%,

respectively) than in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with average rejection rates of 48.7% and 49%, respectively). In addition, most of the model rejection rates in the "mixed" pattern conditions were slightly higher than those in the "all lower" pattern conditions. In contrast, when using a cutoff of 0.90, model rejection rates in the "mixed" and "all lower" pattern conditions were similar. Another inconsistent trend was observed when inspecting the model rejection rates across the three sample size ratios. When using a cutoff value of 0.95, only in the 0.4 loading difference conditions, model rejection rates differed as a function of the sample size ratios (i.e., model rejection rates were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions). When using a cutoff value of 0.90, the trend across the three sample size ratios was observed in both the 0.1 and 0.4 loading difference conditions.

Model Rejection Rates of the TLI

Model rejection rates of the TLI model fit index were investigated under varying conditions in this simulation study. Two TLI cutoff values, 0.90 and 0.95, which were suggested by Bentler and Bonnet (1980) and Hu and Bentler (1999), respectively, were used to evaluate model fit. It means that if the TLI value was less than the relevant cutoff value (0.90 or 0.95) then the null hypothesis of model fit was rejected. Table 22 and Table 23 contain the model rejection rates of the TLI when using cutoff values of 0.90 and 0.95, respectively, in conditions where the true latent mean difference was equal to zero. Table 24 and Table 25 present the model rejection rates of the TLI when using cutoff values of 0.90 and 0.95, respectively, in conditions where the true latent mean difference was equal to 0.5. In each table, values above the dashed line are the model rejection rates of the TLI in the equal factor loading conditions. For the model rejection rates based on the RI strategy, they can be interpreted similarly to Type I error rates

since the estimating models were correctly specified. For the model rejection rates based on the factor-variance scaling method, they can be also interpreted similarly to Type I error rates under the equal factor variance conditions. In addition, model rejection rates can be interpreted similarly to statistical power under the 1.2:0.8 and 0.8:1.2 factor variance ratio conditions since the estimating models were incorrectly specified by constraining unequal factor variances to a value of one across groups. In each table, values below the dashed line are the model rejection rates of the TLI in the unequal factor loading conditions. These rejection rates can be interpreted similarly to statistical power since all the estimating models were incorrectly specified by constraining unequal factor loadings to be equal across groups. In Tables 22 to 25, all model rejection rates that can be interpreted similarly to Type I error rates are italicized.

Latent Mean Difference of Zero

Using a cutoff value of 0.90 in the equal factor loading conditions. Table 22 presents the model rejection rates when using a cutoff of 0.90 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, two out of nine incorrect model rejection rates based on the RI strategy were found to be overly high (8.7% and 9.1%). Both rates were found in the unequal sample size conditions (with a sample size ratio of 1:4 or 4:1). Regarding the model rejection rates based on the factor-variance scaling method, all incorrect rejection rates under the equal factor variance conditions were around 5% and all correct rejection rates under the 0.8:1.2 and 1.2:0.8 factor variance ratio conditions were found to be too low (in the range of 3.8% to 14.1%).

Using a cutoff value of 0.90 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were generally higher in

the 0.4 loading difference conditions than in the 0.1 loading difference conditions. The average model rejection rates were 4.7% and 12.8% for loading difference conditions of 0.1 and 0.4, respectively. However, in a few conditions, opposite trend was observed. For example, in conditions where the sample size ratio was 1:4 and factor loadings were in the "1st loading unequal" or "all lower" pattern, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Across the three factor variance ratios, model rejection rates based on the RI strategy did not differ substantially or systematically. Average model rejection rates were 8.9%, 9% and 8.3% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Regarding the model rejection rates across the four factor loading patterns, clear trends were observed only in conditions in which the loading difference was 0.4. In conditions in which the loading difference was 0.1, model rejection rates in the four loading pattern conditions were all similar. In conditions in which the loading difference was 0.4, model rejection rates in the "mixed" pattern conditions(with a mean of 27.4%) were higher than those in the "1stloading unequal," "2nd loading unequal" and "all lower" pattern conditions (with means of 8.2%, 8.1% and 7.4%, respectively). In addition, model rejection rates in the latter three conditions did not show substantial or systematic differences. Similarly, clear trends across the three sample size ratios were only found in the 0.4 loading difference conditions. Specifically, model rejection rates were generally lower in the 1:4 sample size ratio conditions than in the 1:1 and 4:1 sample size ratio conditions. In addition, model rejection rates were similar in the 1:1 and 4:1 conditions. Within the 0.4 loading difference conditions, average model rejection rates were 16.3%, 15.6% and 6.4%, respectively, for the 1:1, 4:1 and 1:4 sample size ratio conditions.

Table 22

Model Rejection Rates of the TLI When Using a Cutoff of 0.90 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.060*	*0.058*	*0.038*	0.089	*0.046*	0.095
		100:400	*0.051*	*0.043*	*0.087*	0.128	*0.030*	0.038
		400:100	*0.060*	*0.052*	*0.031*	0.045	*0.091*	0.141
0.1	1st Loading Unequal	250:250	0.055	0.052	0.051	0.066	0.038	0.116
		100:400	0.044	0.045	0.068	0.086	0.014	0.042
		400:100	0.048	0.050	0.029	0.039	0.094	0.149
	2nd Loading Unequal	250:250	0.048	0.049	0.044	0.060	0.037	0.109
		100:400	0.046	0.044	0.062	0.071	0.025	0.048
		400:100	0.048	0.050	0.036	0.047	0.083	0.162
	All Lower	250:250	0.051	0.056	0.036	0.052	0.036	0.142
		100:400	0.034	0.035	0.043	0.059	0.014	0.042
		400:100	0.068	0.072	0.035	0.040	0.082	0.176
	Mixed	250:250	0.047	0.046	0.048	0.097	0.052	0.094
		100:400	0.052	0.052	0.061	0.092	0.027	0.053
		400:100	0.048	0.039	0.022	0.044	0.063	0.109
0.4	1st Loading Unequal	250:250	0.090	0.176	0.120	0.144	0.054	0.481
		100:400	0.019	0.031	0.048	0.044	0.006	0.058
		400:100	0.157	0.206	0.089	0.076	0.159	0.474
	2nd Loading Unequal	250:250	0.089	0.183	0.114	0.103	0.059	0.477
		100:400	0.023	0.047	0.069	0.060	0.003	0.064
		400:100	0.116	0.175	0.086	0.078	0.172	0.503
	All Lower	250:250	0.036	0.352	0.085	0.142	0.004	0.710
		100:400	0.000	0.009	0.000	0.001	0.001	0.009
		400:100	0.180	0.562	0.189	0.240	0.167	0.899
	Mixed	250:250	0.478	0.460	0.426	0.543	0.403	0.541
		100:400	0.188	0.187	0.352	0.375	0.057	0.113
		400:100	0.171	0.171	0.054	0.115	0.334	0.357

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Table 22 also contains the model rejection rates based on the factor-variance scaling method when using a cutoff of 0.90 in conditions in which the latent mean difference was equal to zero. In the unequal factor loading conditions, model rejection rates based on the factor-

variance scaling method were slightly higher than those based on the RI strategy. In addition, the trends in the rejection rates based on the factor-variance scaling method were consistent with those based on the RI strategy. First, model rejection rates based on the factor-variance scaling method increased as the loading difference magnitude increased. For the loading difference conditions of 0.1 and 0.4, average rejection rates were 7.2% and 25.5%, respectively. Second, the trends in the model rejection rates across the four factor loading patterns were different for the 0.1 and 0.4 loading difference conditions. In conditions in which the loading difference was 0.1, model rejection rates based on the factor-variance scaling method did not vary greatly as a function of the factor loading patterns. In conditions in which the loading difference was 0.4, model rejection rates based on the factor-variance scaling method were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions with the same average rejection rate of 18.8%. Higher rejection rates were found in the "all lower" and "mixed" pattern conditions with means of 32.5% and 31.8%, respectively. Third, model rejection rates based on the factor-variance scaling method were generally higher in conditions in which the factor variance ratio was 0.8:1.2 than in conditions in which the factor variance ratio was 1:1 or 1.2:0.8. This trend was most obvious in conditions with a loading difference of 0.4 in which more than half of the model rejection rates in the 0.8:1.2 facto variance ratio conditions were above 40% whereas more than half of the model rejection rates in the 1:1 and 1.2:0.8 factor variance ratio conditions were below 20%. Average model rejection rates were 13.1%, 11.1% and 24.7% for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively. Last, the 1:1 and 4:1 sample size ratio conditions led to higher model rejection rates than did the 1:4 sample size ratio conditions, particularly within conditions with a loading difference of 0.4. Additionally, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. The average model

rejection rates were 35.9%, 32.1% and 8.3% for the respective sample size ratio conditions of 1:1, 4:1 and 1:4.

Using a cutoff value of 0.95 in the equal factor loading conditions. Table 23 presents the model rejection rates of the TLI when using a cutoff value of 0.95 in conditions where the true latent mean difference was equal to zero. Compared to the model rejection rates in Table 22, it was found that using a cutoff value of 0.95 led to higher model rejection rates than did using a cutoff value of 0.90. In the equal factor loading conditions, model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, differed substantially from 5% (ranging from 15.4% to 24.9%). For the model rejection rates based on the factor-variance scaling method, they were overly high in the equal factor variance conditions (in the range of 18.5% to 21.1%). In the unequal factor variance conditions, model rejection rates based on the factor-variance scaling method, which can be interpreted similarly to statistical power, were low (ranging from 21.1% to 31%).

Using a cutoff value of 0.95 in the unequal factor loading conditions. When using a cutoff value of 0.95, model rejection rates in the unequal factor loading conditions were higher than those observed when using a cutoff value of 0.90. Additionally, using a cutoff value of 0.95, the trends across loading difference magnitudes, factor loading patterns, factor variance ratios and sample size ratios were consistent with those found when using a cutoff value of 0.90.

Table 23

Model Rejection Rates of the TLI When Using a Cutoff of 0.95 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.216*	*0.211*	*0.179*	0.298	*0.187*	0.287
		100:400	*0.198*	*0.191*	*0.249*	0.310	*0.164*	0.212
		400:100	*0.194*	*0.185*	*0.154*	0.211	*0.232*	0.301
0.1	1st Loading Unequal	250:250	0.187	0.189	0.202	0.270	0.187	0.351
		100:400	0.174	0.178	0.232	0.276	0.138	0.227
		400:100	0.200	0.196	0.167	0.205	0.247	0.361
	2nd Loading Unequal	250:250	0.181	0.173	0.180	0.236	0.181	0.357
		100:400	0.186	0.195	0.228	0.253	0.156	0.236
		400:100	0.208	0.211	0.153	0.197	0.261	0.385
	All Lower	250:250	0.203	0.222	0.190	0.217	0.175	0.404
		100:400	0.173	0.185	0.188	0.206	0.121	0.242
		400:100	0.208	0.212	0.173	0.195	0.232	0.406
	Mixed	250:250	0.203	0.205	0.195	0.321	0.201	0.311
		100:400	0.203	0.194	0.220	0.284	0.168	0.226
		400:100	0.179	0.183	0.144	0.205	0.222	0.300
0.4	1st Loading Unequal	250:250	0.339	0.529	0.380	0.367	0.298	0.835
		100:400	0.187	0.293	0.265	0.261	0.096	0.399
		400:100	0.405	0.492	0.324	0.308	0.410	0.779
	2nd Loading Unequal	250:250	0.342	0.532	0.404	0.399	0.285	0.831
		100:400	0.173	0.293	0.276	0.273	0.098	0.408
		400:100	0.390	0.473	0.304	0.303	0.441	0.790
	All Lower	250:250	0.311	0.847	0.446	0.593	0.161	0.980
		100:400	0.045	0.225	0.117	0.181	0.018	0.398
		400:100	0.550	0.865	0.553	0.616	0.519	0.982
	Mixed	250:250	0.868	0.859	0.821	0.899	0.826	0.895
		100:400	0.607	0.610	0.756	0.777	0.384	0.567
		400:100	0.614	0.622	0.391	0.556	0.753	0.765

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

When the RI strategy was implemented, model rejection rates were generally higher in the 0.4 loading difference conditions (with a mean of 39.3%) than in the 0.1 loading difference conditions (with a mean of 19.1%). Similar to the rejection rates observed when using a cutoff

value of 0.90, there were some exceptions when using a cutoff value of 0.95. For example, in conditions in which the sample size ratio was 1:4, the factor variance ratio was 0.8:1.2 and factor loadings were in the "1st loading unequal," "2nd loading unequal" or "all lower" pattern conditions, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Across the three factor variance ratio conditions, model rejection rates did not show substantial or systematic differences. Average rejection rates were 29.7%, 30.5% and 27.4% for the respective 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions. Across the four factor loading patterns, clear trends were only found in the 0.4 loading difference conditions. Within conditions with a loading difference of 0.4, model rejection rates were similar in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions with average rejection rates of 30%, 30.1% and 30.2%, respectively. Much higher model rejection rates were found in the "mixed" pattern conditions with an average rejection rate of 66.9%. For the three sample size ratios, the 1:1 and 4:1 sample size ratio conditions generally led to higher model rejection rates than did the 1:4 sample size ratio conditions, particularly in conditions in which the loading difference was 0.4. Additionally, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. Within the 0.4 loading difference conditions, average model rejection rates were 45.7%, 47.1% and 25.2% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

When the factor-variance scaling method was used, model rejection rates observed when using a cutoff of 0.95 were also higher than those observed when using a cutoff of 0.90. Additionally, the trends across loading difference magnitudes, factor loading patterns and factor variance ratios were consistent with those obtained when using a cutoff of 0.90. First, model rejection rates based on the factor-variance scaling method were higher when the loading

difference magnitude was larger. Average rejection rates were 25% and 57.8% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratios, the 0.8:1.2 factor variance ratio conditions led to higher model rejection rates than did the 1:1 and 1.2:0.8 factor variance ratio conditions. Several high rejection rates (e.g., 98% and 98.2%) were only observed in the 0.8:1.2 conditions. Average model rejection rates were 51.8%, 37.4% and 35%, respectively, for the 0.8:1.2, 1:1 and 1.2:0.8 factor variance ratio conditions. Inspecting the model rejection rates across the four factor loading patterns, it was found that within the 0.1 loading difference conditions, model rejection rates did not differ substantially or systematically as a function of the factor loading patterns. Model rejection rates, however, varied across the four factor loading patterns in the 0.4 loading difference conditions. More specifically, model rejection rates were higher in the "all lower" and "mixed" pattern conditions (with average rejection rates of 63.2% and 72.8%, respectively) than in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with average rejection rates of 47.4% and 47.8%, respectively). Across the three sample size ratios, model rejection rates did not show clear trend in conditions in which the loading difference was 0.1. However, in conditions in which the loading difference was 0.4, model rejection rates were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. In addition, model rejection rates were generally higher in the 1:1 sample size ratio conditions than in the 4:1 sample size ratio conditions. This trend was slightly different from that observed when using a cutoff of 0.90 in which model rejection rates in the 1:1 and 4:1 conditions did not differ substantially. Within conditions with a loading difference of 0.4, average model rejection rates were 71.4%, 62.9% and 39.0% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

Latent Mean Difference of 0.5

Using a cutoff value of 0.90 in the equal factor loading conditions. Table 24 contains the model rejection rates of the TLI when using a cutoff value of 0.90 in conditions where the true latent mean difference was equal to 0.5. In the equal factor loading conditions, one incorrect model rejection rate based on the RI strategy was found to be overly high (8%). When the factor-variance scaling method was implemented, all incorrect model rejection rates in the equal factor variance conditions were around 5%, and all correct model rejection rates in the unequal factor variance conditions (with a factor variance ratio of 1.2:0.8 or 0.8:1.2) were found to be overly low (ranging from 4.9% to 12%).

Using a cutoff value of 0.90 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were investigated while varying the loading difference magnitude, factor variance ratio, factor loading pattern and sample size ratio. Across the two loading difference magnitudes, model rejection rates were generally higher when the loading difference magnitude was larger. Within conditions with a loading difference of 0.1, model rejection rates were all below 10% and the average rejection rate was 4.9%. Within conditions with a loading difference of 0.4, many model rejection rates were greater than 10% with the highest rejection rate of 59.8%, and the average rate was 16.4%. However, in a few conditions, opposite trend was observed. For example, in conditions in which the sample size ratio was 1:4 and factor loadings were in the "1st loading unequal" or "all lower" pattern conditions, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Two zero rejection rates were found in the 0.4 loading difference conditions. Across the three factor variance ratios, model rejection rates did not show substantial differences and average model rejection rates were 10.7%, 11.4% and 9.9% for the

factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Across the four factor loading patterns, different trends were observed in the 0.1 and 0.4 loading difference conditions. In conditions in which the loading difference was 0.1, model rejection rates in the four factor loading pattern conditions were similar. In conditions in which the loading difference was 0.4, model rejection rates were higher in the "mixed" pattern conditions than in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. Additionally, model rejection rates in the latter three conditions did not differ substantially or systematically. Average rejection rates were 9.9%, 9.6%, 9.1% and 36.9% for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions, respectively. The trends across the three sample size ratios were also more obvious in the 0.4 loading difference conditions. Specifically, the sample size ratio conditions of 1:1 and 4:1 consistently led to higher model rejection rates than did the sample size ratio condition of 1:4. In addition, the sample size ratio condition of 4:1 generally led to higher rejection rates than did the sample size ratio condition of 1:1, although the differences in model rejection rates were not large. Within conditions with a loading difference of 0.4, average model rejection rates were 19.4%, 23.0% and 6.8% for the 1:1, 4:1 and 1:4 sample size ratio conditions, respectively.

When using a cutoff value of 0.90 in conditions in which the latent mean difference was equal to 0.5, model rejection rates when the factor-variance scaling method was used were also investigated. For the two loading difference magnitudes, model rejection rates based on the factor-variance scaling method were generally higher when the loading difference magnitude was larger. In conditions in which the loading difference was 0.1, model rejection rates were in the range of 2.8% to 17.2% with a mean of 7.5%. In conditions in which the loading difference was 0.4, most of the model rejection rates were greater than 10% with a mean of 29.6%.

Table 24

Model Rejection Rates of the TLI When Using a Cutoff of 0.90 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.046*	*0.042*	*0.060*	0.104	*0.048*	0.085
		100:400	*0.056*	*0.051*	*0.073*	0.116	*0.027*	0.056
		400:100	*0.067*	*0.061*	*0.028*	0.049	*0.080*	0.120
0.1	1st Loading Unequal	250:250	0.048	0.055	0.050	0.080	0.040	0.117
		100:400	0.038	0.048	0.074	0.091	0.032	0.055
		400:100	0.057	0.059	0.034	0.051	0.088	0.159
	2nd Loading Unequal	250:250	0.054	0.057	0.057	0.082	0.043	0.118
		100:400	0.034	0.034	0.062	0.091	0.029	0.047
		400:100	0.053	0.051	0.034	0.050	0.077	0.146
	All Lower	250:250	0.046	0.053	0.056	0.065	0.038	0.138
		100:400	0.025	0.028	0.049	0.054	0.017	0.035
		400:100	0.042	0.042	0.035	0.046	0.084	0.172
	Mixed	250:250	0.048	0.042	0.049	0.093	0.051	0.098
		100:400	0.049	0.051	0.080	0.110	0.031	0.051
		400:100	0.056	0.056	0.020	0.035	0.087	0.127
0.4	1st Loading Unequal	250:250	0.085	0.192	0.119	0.112	0.067	0.489
		100:400	0.020	0.047	0.053	0.051	0.009	0.069
		400:100	0.192	0.265	0.144	0.133	0.198	0.556
	2nd Loading Unequal	250:250	0.095	0.202	0.144	0.136	0.047	0.482
		100:400	0.012	0.027	0.067	0.063	0.001	0.064
		400:100	0.174	0.248	0.117	0.117	0.210	0.564
	All Lower	250:250	0.037	0.338	0.086	0.155	0.014	0.728
		100:400	0.000	0.008	0.004	0.006	0.000	0.011
		400:100	0.233	0.616	0.246	0.286	0.195	0.928
	Mixed	250:250	0.598	0.587	0.595	0.693	0.441	0.590
		100:400	0.198	0.199	0.386	0.400	0.061	0.129
		400:100	0.363	0.343	0.169	0.276	0.514	0.538

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Across the three factor variance ratios, model rejection rates based on the factor-variance scaling method were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. This trend was most obvious in conditions in

which the loading difference was 0.4. Average model rejection rates were 26.7%, 15.2% and 13.7%, respectively for the 0.8:1.2, 1:1 and 1.2:0.8 factor variance ratio conditions. Inspecting the model rejection rates across the four factor loading pattern conditions, clear trends were only observed in conditions in which the loading difference was 0.4. In conditions in which the loading difference of 0.1, model rejection rates were similar in the four loading pattern conditions with average rejection rates of 7.9%, 7.5%, 7% and 7.4%, respectively. In conditions in which the loading difference was 0.4, model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions with average rejection rates of 21.3% and 21.1%, respectively. The "mixed" pattern conditions consistently led to higher model rejection rates than did the "1st loading unequal" and "2nd loading unequal" pattern conditions. The "all lower" pattern conditions led to higher rejection rates than did the "1st loading unequal" and "2nd loading unequal" pattern conditions in conditions in which the sample size ratio was 1:1 or 4:1. Average rejection rates were 34.2% and 41.7% for the "all lower" and "mixed" pattern conditions, respectively. Similarly, clear trends across the three sample size ratios were only found in the 0.4 loading difference conditions. The 1:1 and 4:1 sample size ratio conditions generally produced higher model rejection rates than did the 1:4 sample size ratio conditions. In addition, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. Within conditions with a loading difference of 0.4, average rejection rates were 39.2%, 40.6% and 9.0%, respectively, for the 1:1, 4:1 and 1:4 sample size ratio conditions.

Using a cutoff value of 0.95 in the equal factor loading conditions. Table 25 contains the model rejection rates of the TLI when using a cutoff value of 0.95 in conditions where the true latent mean difference was equal to 0.5. Compared to the model rejection rates in Table 24 in which a cutoff value of 0.90 was used, it is obvious that using a cutoff of 0.95 led to higher

model rejection rates than did using a cutoff of 0.90, regardless of the factor scaling method used. In the equal factor loading conditions, the incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were much higher than 5% (in the range of 15.4% to 23.2%). When the factor-variance scaling method was used, the model rejection rates under the equal factor variance conditions, which can also be interpreted similarly to Type I error rates, were found to be overly high (in the range of 18.3% to 21.3%). The model rejection rates under the unequal factor variance conditions, which can be interpreted similarly to statistical power, were found to be low (ranging from 21.9% to 31.7%).

Using a cutoff value of 0.95 in the unequal factor loading conditions. When using a cutoff of 0.95 in the unequal factor loading conditions, the trends across loading difference magnitude, factor variance ratios, factor loading patterns and sample size ratios were consistent with those found when using a cutoff value of 0.90. First, model rejection rates when the RI strategy was used were generally higher in the 0.4 loading difference conditions (with a mean of 43.6%) than in the 0.1 loading difference conditions (with a mean of 19.5%). However, opposite trends were observed in a few conditions. For example, in conditions in which the sample size ratio was 1:4 and factor loadings were in the "1st loading unequal," "2nd loading unequal" or "all lower" pattern conditions, model rejection rates were lower in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Second, model rejection rates did not differ substantially or systematically as a function of the factor variance ratios. Average model rejection rates were 34.2%, 31.8% and 28.7% for the factor variance ratio conditions of 1.2:0.8, 1:1 and 0.8:1.2, respectively.

Table 25

Model Rejection Rates of the TLI When Using a Cutoff of 0.95 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.193*	*0.196*	*0.197*	0.309	*0.190*	0.297
		100:400	*0.196*	*0.183*	*0.227*	0.288	*0.168*	0.219
		400:100	*0.230*	*0.213*	*0.154*	0.232	*0.232*	0.317
0.1	1st Loading Unequal	250:250	0.197	0.200	0.199	0.264	0.179	0.337
		100:400	0.204	0.208	0.235	0.263	0.164	0.253
		400:100	0.213	0.219	0.171	0.206	0.237	0.352
	2nd Loading Unequal	250:250	0.202	0.206	0.214	0.271	0.181	0.356
		100:400	0.167	0.166	0.222	0.269	0.140	0.222
		400:100	0.200	0.203	0.175	0.222	0.238	0.363
	All Lower	250:250	0.194	0.214	0.210	0.241	0.169	0.412
		100:400	0.166	0.172	0.221	0.247	0.118	0.239
		400:100	0.192	0.204	0.168	0.183	0.246	0.419
	Mixed	250:250	0.208	0.208	0.203	0.320	0.201	0.316
		100:400	0.186	0.183	0.232	0.302	0.151	0.219
		400:100	0.216	0.231	0.156	0.215	0.262	0.338
0.4	1st Loading Unequal	250:250	0.362	0.539	0.450	0.438	0.310	0.829
		100:400	0.161	0.261	0.278	0.266	0.093	0.397
		400:100	0.461	0.554	0.422	0.418	0.480	0.817
	2nd Loading Unequal	250:250	0.378	0.566	0.449	0.442	0.268	0.827
		100:400	0.164	0.271	0.275	0.258	0.088	0.410
		400:100	0.461	0.547	0.387	0.378	0.504	0.825
	All Lower	250:250	0.322	0.832	0.443	0.583	0.178	0.980
		100:400	0.049	0.239	0.121	0.175	0.009	0.404
		400:100	0.617	0.898	0.624	0.678	0.528	0.993
	Mixed	250:250	0.927	0.917	0.919	0.953	0.864	0.952
		100:400	0.626	0.633	0.774	0.783	0.414	0.578
		400:100	0.764	0.760	0.656	0.778	0.869	0.882

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Third, clear trends across the four factor loading patterns were found only in conditions in which the loading difference was 0.4. Specifically, model rejection rates were similar in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions with average

rejection rates of 33.5%, 33% and 32.1%, respectively. However, model rejection rates were higher in the "mixed" pattern conditions than in the other three pattern conditions. All model rejection rates in the "mixed" pattern conditions were greater than 40% and two of them were above 90%. The average rejection rate for the "mixed" pattern conditions was 75.7%. Finally, model rejection rates were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. This trend was most obvious within conditions with a loading difference of 0.4, in which average rejection rates were 48.9%, 56.4% and 25.4% for the respective 1:1, 4:1 and 1:4 sample size ratio conditions.

Table 25 also contains the model rejection rates based on the factor-variance scaling method when using a cutoff value of 0.95 in conditions in which the true latent mean difference was equal to 0.5. When the factor-variance scaling method was implemented, model rejection rates when using a cutoff value of 0.95 were higher than those observed when using a cutoff value of 0.90. In addition, when using a cutoff value of 0.95, the trends across loading difference magnitudes, factor variance ratios, factor loading patterns and sample size ratios were consistent with those found when using a cutoff value of 0.90. More specifically, model rejection rates based on the factor-variance scaling method increased as the loading difference magnitude increased. Within conditions with a loading difference of 0.1, most of the model rejection rates were between 20% and 35% and the average rejection rate was 25.7%. Within conditions with a loading difference of 0.4, most of the model rejection rates exceeded 40% and several of them were greater than 90%. The average rejection rate was 61.3% for the 0.4 loading difference conditions. For the three factor variance ratio conditions, model rejection rates were generally higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions. This trend was most obvious in the 0.4 loading difference conditions in which

two thirds of the model rejection rates under the 0.8:1.2 factor variance ratio conditions were above 80%. Additionally, model rejection rates were similar in the 1:1 and 1.2:0.8 factor variance ratio conditions. Average rejection rates were 39.3%, 38.1% and 53% for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively. Regarding the model rejection rates across the four factor loading patterns, no clear trend was found when the loading difference was equal to 0.1. However, within conditions with a loading difference of 0.4, model rejection rates differed as a function of the factor loading patterns. Specifically, the "mixed" pattern conditions generally led to higher model rejection rates than did the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions, and the latter three pattern conditions led to similar rejection rates. The average rejection rates were 80.4%, 50.2% and 50.3% and 64.2% for the "mixed," "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions, respectively. Similarly, clear trends across the three sample size ratio conditions were found only in conditions in which the loading difference was 0.4. Model rejection rates were generally higher in the 1:1 and 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions. Additionally, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 73.8%, 71.1% and 39.0% for the respective 1:1, 4:1 and 1:4 sample size ratio conditions.

Model Rejection Rates of the SRMR

In the current study, the performance of the SRMR model fit index in terms of the correct and incorrect model rejection rates was investigated under varying conditions. Two SRMR cutoff values, 0.05 and 0.08, which were proposed by Steiger (1989) and Hu and Bentler (1999), respectively, were used to determine whether the null hypothesis of model fit should be rejected. If the SRMR was greater than the relevant cutoff value (0.05 or 0.08) then the null hypothesis of

model fit was rejected. Table 26 and Table 27 respectively present the model rejection rates of the SRMR using cutoff value of 0.05 and 0.08 in conditions where the true latent mean difference was equal to zero. Table 28 and Table 29 contain the model rejection rates of the SRMR when using cutoff values of 0.05 and 0.08, respectively, in conditions where the true latent mean difference was equal to 0.5. In each table, values above the dashed line are the model rejection rates in the equal factor loading conditions. When the RI strategy was implemented, the estimating models were correctly specified. Thus, the model rejection rates of the SRMR can be interpreted similarly to Type I error rates. When the factor-variance scaling method was used, the model rejection rates of the SRMR in the equal factor variance conditions can also be interpreted similarly to Type I error rates. The model rejection rates in the unequal factor variance ratio conditions, on the other hand, can be interpreted similarly to statistical power since the estimating models were incorrectly specified by constraining unequal factor variances to a value of one across groups. In each table, values below the dashed line are the model rejection rates in the unequal factor loading conditions. Since the estimating models were incorrectly specified by constraining unequal factor loadings to be equal across groups, model rejection rates in the unequal factor loading conditions can be interpreted similarly to statistical power. In Tables 26 to 29, all incorrect rejection rates that can be interpreted similarly to Type I error rates are italicized.

Latent Mean Difference of Zero

Using a cutoff value of 0.05 in the equal factor loading conditions. Table 26 contains the model rejection rates of the SRMR when using a cutoff value of 0.05 in conditions where the true latent mean difference was equal to zero.

Table 26

Model Rejection Rates of the SRMR When using a Cutoff of 0.05 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.093*	*0.200*	*0.068*	0.471	*0.076*	0.474
		100:400	*0.086*	*0.171*	*0.088*	0.340	*0.073*	0.398
		400:100	*0.091*	*0.182*	*0.081*	0.418	*0.084*	0.336
0.1	1st Loading Unequal	250:250	0.095	0.203	0.099	0.410	0.095	0.622
		100:400	0.097	0.230	0.105	0.302	0.087	0.520
		400:100	0.090	0.189	0.078	0.369	0.109	0.447
	2nd Loading Unequal	250:250	0.084	0.200	0.090	0.375	0.089	0.656
		100:400	0.097	0.220	0.087	0.297	0.108	0.550
		400:100	0.081	0.199	0.093	0.371	0.101	0.456
	All Lower	250:250	0.122	0.299	0.107	0.339	0.103	0.746
		100:400	0.098	0.278	0.085	0.243	0.089	0.650
		400:100	0.112	0.226	0.104	0.322	0.096	0.536
	Mixed	250:250	0.127	0.246	0.113	0.562	0.132	0.554
		100:400	0.121	0.227	0.104	0.374	0.117	0.483
		400:100	0.112	0.220	0.096	0.483	0.099	0.388
0.4	1st Loading Unequal	250:250	0.379	0.778	0.395	0.516	0.321	0.984
		100:400	0.242	0.691	0.290	0.424	0.163	0.933
		400:100	0.338	0.566	0.289	0.423	0.315	0.865
	2nd Loading Unequal	250:250	0.384	0.791	0.406	0.537	0.326	0.981
		100:400	0.225	0.654	0.294	0.424	0.165	0.931
		400:100	0.306	0.547	0.269	0.421	0.306	0.883
	All Lower	250:250	0.625	0.998	0.695	0.908	0.437	1.000
		100:400	0.296	0.971	0.376	0.721	0.201	1.000
		400:100	0.666	0.972	0.651	0.802	0.578	0.999
	Mixed	250:250	0.935	0.961	0.900	0.989	0.904	0.980
		100:400	0.741	0.834	0.816	0.903	0.583	0.952
		400:100	0.728	0.840	0.595	0.948	0.814	0.909

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

In the equal factor loading conditions, seven out of nine incorrect model rejection rates based on the RI strategy differed substantially from 5%. In addition, all three incorrect model rejection rates based on the factor-variance scaling method were much higher than 5% (in the

range of 17.1% to 20%). All six correct model rejection rates based on the factor-variance scaling method were found to be low (ranging from 33.6% to 47.4%).

Using a cutoff value of 0.05 in the unequal factor loading conditions. When using a cutoff of 0.05 in the unequal factor loading conditions, model rejection rates were examined across loading difference magnitudes, factor variance ratios, factor loading patterns and sample size ratios. First, model rejection rates based on the RI strategy were higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.4, model rejection rates were in the range of 16.3% to 93.5% and several of them were greater than 90%. In contrast, within conditions with a loading difference of 0.1, model rejection rates were in the range of 7.8% to 13.2%. Average model rejection rates were 47.1% and 10.1% for the loading difference conditions of 0.4 and 0.1, respectively. Second, model rejection rates based on the RI strategy did not vary substantially or systematically as a function of the factor variance ratios. For the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, average model rejection rates were 29.6%, 29.7% and 26.4%, respectively. Third, model rejection rates varied systematically across the four factor loading patterns, particularly within conditions with a loading difference of 0.4. More specifically, the "mixed" pattern consistently led to higher model rejection rates than did the "1st loading unequal," "2nd loading unequal" and "all lower" patterns. The "all lower" pattern also led to higher model rejection rates than did the "1st loading unequal" and "2nd loading unequal" patterns. In addition, model rejection rates were similar in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 30.4%, 29.8%, 50.3% and 78%, respectively, for the "1st loading unequal," "2nd loading unequal,""all lower" and "mixed" pattern conditions. Last, no clear trend across the three sample size ratios was found

in conditions in which the loading difference was 0.1. However, in conditions in which the loading difference was 0.4, model rejection rates in the 1:1 sample size ratio conditions were generally higher than those in the 1:4 or 4:1 sample size ratio conditions. In addition, most of the model rejection rates in the 4:1 sample size ratio conditions were higher than those in the 1:4 sample size ratio conditions. For the 1:1, 4:1 and 1:4 sample size ratio conditions, again within conditions with a loading difference of 0.4, average model rejection rates were 55.9%, 48.8% and 36.6%, respectively.

Table 26 also contains the model rejection rates of the SRMR when implementing the factor-variance scaling method. In the unequal factor loading conditions, model rejection rates based on the factor-variance scaling method were higher when the magnitude of the loading difference was larger. In conditions in which the loading difference was 0.1, model rejection rates were in the range of 18.9% to 74.6% with a mean of 38.3%. On the other hand, in conditions in which the loading difference was 0.4, half of the model rejection rates were greater than 90% and two of them were equal to 100%. The average model rejection rate for the 0.4 loading difference conditions was 80.7%. Across the three factor variance ratio conditions, model rejection rates based on the factor-variance scaling method were higher in the 0.8:1.2 factor variance ratio conditions (with an average rejection rate of 75.1%) than in the 1:1 and 1.2:0.8 conditions (with average rejection rates of 51.4% and 51.9%, respectively). For the four factor loading patterns, clear trends were only observed in conditions in which the loading difference was 0.4. Specifically, model rejection rates were generally higher in the "all lower" and "mixed" pattern conditions than in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 93%, 92.4%, 68.7% and 68.5% for the "all lower," "mixed," "1st loading unequal" and

"2nd loading unequal" pattern conditions, respectively. Similarly, the sample size ratio only produced clear trends in the 0.4 loading difference conditions. Specifically, model rejection rates were generally higher in the equal sample size conditions than in the unequal sample size conditions (with a sample size ratio of 1:4 or 4:1). In addition, model rejection rates were similar in the two unequal sample size conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 86.9%, 78.7% and 76.5% for the respective sample size ratio conditions of 1:1, 1:4 and 4:1.

Using a cutoff value of 0.08 in the equal factor loading conditions. Table 27 presents the model rejection rates of the SRMR when using a cutoff value of 0.08 in conditions where the true latent mean difference was equal to zero. In the equal factor loading conditions, the model rejection rates obtained when using a cutoff of 0.08 were lower than those in Table 26 in which a cutoff value of 0.05 was used. When the RI strategy was used, all incorrect model rejection rates were equal to zero. When the factor-variance scaling method was implemented, two out of three incorrect model rejection rates were equal to zero and the third one was equal to 0.1%. In addition, all correct model rejection rates that were based on the factor-variance scaling method were found to be extremely low (in the range of 0% to 0.4%).

Using a cutoff value of 0.08 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates observed when using a cutoff value of 0.08 were much lower than those obtained when using a cutoff value of 0.05. When implementing the RI strategy, most of the model rejection rates were equal to zero. A few non-zero rejection rates were observed in conditions in which the loading difference was 0.4 and factor loadings were in the "all lower" or "mixed" pattern conditions. These non-zero rejection rates had very low values with a highest rejection rate of 2%.

Table 27

Model Rejection Rates of the SRMR When Using a Cutoff of 0.08 in Conditions where the Latent Mean Difference is Zero

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.000*	*0.001*	*0.000*	0.004	*0.000*	0.004
		100:400	*0.000*	*0.000*	*0.000*	0.001	*0.000*	0.000
		400:100	*0.000*	*0.000*	*0.000*	0.001	*0.000*	0.000
0.1	1st Loading Unequal	250:250	0.000	0.000	0.000	0.000	0.000	0.003
		100:400	0.000	0.000	0.000	0.000	0.000	0.002
		400:100	0.000	0.000	0.000	0.000	0.000	0.003
	2nd Loading Unequal	250:250	0.000	0.000	0.000	0.001	0.000	0.005
		100:400	0.000	0.000	0.000	0.000	0.000	0.000
		400:100	0.000	0.000	0.000	0.000	0.000	0.001
	All Lower	250:250	0.000	0.000	0.000	0.001	0.000	0.028
		100:400	0.000	0.000	0.000	0.000	0.000	0.010
		400:100	0.000	0.000	0.000	0.001	0.000	0.007
	Mixed	250:250	0.000	0.001	0.000	0.005	0.000	0.003
		100:400	0.000	0.000	0.000	0.001	0.000	0.007
		400:100	0.000	0.000	0.000	0.004	0.000	0.001
0.4	1st Loading Unequal	250:250	0.000	0.011	0.000	0.000	0.000	0.329
		100:400	0.000	0.017	0.000	0.000	0.000	0.178
		400:100	0.000	0.002	0.000	0.000	0.000	0.048
	2nd Loading Unequal	250:250	0.000	0.028	0.000	0.000	0.000	0.324
		100:400	0.000	0.010	0.000	0.000	0.000	0.207
		400:100	0.000	0.000	0.000	0.000	0.000	0.046
	All Lower	250:250	0.000	0.544	0.001	0.035	0.000	0.973
		100:400	0.000	0.253	0.000	0.020	0.000	0.828
		400:100	0.001	0.137	0.001	0.006	0.000	0.605
	Mixed	250:250	0.020	0.042	0.011	0.188	0.010	0.186
		100:400	0.001	0.014	0.006	0.026	0.000	0.169
		400:100	0.002	0.018	0.000	0.166	0.008	0.021

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

When the factor-variance scaling method was implemented, model rejection rates were slightly higher than those based on the RI strategy although they were still low. Within conditions with a loading difference of 0.1, half of the model rejection rates were equal to zero.

In addition, most of the non-zero rejection rates were lower than 1%. Within conditions with a loading difference of 0.4, model rejection rates increased. Only seven out of thirty-six model rejection rates were equal to zero. Average model rejection rates were 0.2% and 15.1% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratio conditions, model rejection rates were higher under the 0.8:1.2 factor variance ratio conditions than under the 1:1 and 1.2:0.8 factor variance ratio conditions. In conditions in which the factor variance ratio was 0.8:1.2, only one model rejection rate was equal to zero. Additionally, a few high and relatively high rejection rates (e.g., 97.3% and 82.8%) were only found in the 0.8:1.2 factor variance ratio conditions. Average model rejection rates were 4.5%, 1.9% and 16.6% for the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, respectively. Regarding the model rejection rates across the four factor loading patterns, clear trends were only found in conditions in which the loading difference was 0.4. More specifically, model rejection rates were consistently higher in the "all lower" pattern conditions than in the "1st loading unequal" and "2nd loading unequal" pattern conditions. In addition, in conditions in which the factor variance ratio was 1:1 or 0.8:1.2, model rejection rates were higher in the "all lower" pattern conditions than in the "mixed" pattern conditions. Across the three sample size ratios, clear trends were also found in the 0.4 loading difference conditions. Specifically, the sample size ratio condition of 1:1 generally led to higher model rejection rates than did the sample size ratio conditions of 1:4 and 4:1. For the 1:1, 1:4 and 4:1 sample size ratio conditions, average model rejection rates were 22.2%, 14.4% and 8.7%, respectively.

Table 28

Model Rejection Rates of the SRMR When Using a Cutoff of 0.05 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.085*	*0.189*	*0.104*	0.501	*0.088*	0.484
		100:400	*0.089*	*0.187*	*0.091*	0.329	*0.091*	0.416
		400:100	*0.118*	*0.208*	*0.086*	0.429	*0.098*	0.372
0.1	1st Loading Unequal	250:250	0.099	0.236	0.108	0.416	0.093	0.609
		100:400	0.119	0.246	0.108	0.300	0.112	0.556
		400:100	0.119	0.234	0.108	0.374	0.106	0.476
	2nd Loading Unequal	250:250	0.117	0.240	0.121	0.415	0.104	0.646
		100:400	0.088	0.212	0.096	0.312	0.103	0.564
		400:100	0.102	0.200	0.101	0.376	0.106	0.452
	All Lower	250:250	0.119	0.302	0.135	0.372	0.105	0.770
		100:400	0.103	0.277	0.110	0.270	0.099	0.648
		400:100	0.104	0.219	0.098	0.302	0.121	0.585
	Mixed	250:250	0.137	0.255	0.131	0.587	0.146	0.580
		100:400	0.128	0.218	0.121	0.384	0.108	0.494
		400:100	0.132	0.267	0.111	0.490	0.133	0.436
0.4	1st Loading Unequal	250:250	0.418	0.792	0.472	0.602	0.371	0.985
		100:400	0.231	0.682	0.297	0.437	0.177	0.954
		400:100	0.381	0.620	0.377	0.509	0.369	0.918
	2nd Loading Unequal	250:250	0.432	0.799	0.464	0.568	0.311	0.981
		100:400	0.226	0.686	0.305	0.447	0.183	0.941
		400:100	0.387	0.624	0.326	0.467	0.367	0.907
	All Lower	250:250	0.619	0.997	0.708	0.920	0.441	1.000
		100:400	0.299	0.964	0.384	0.740	0.192	0.999
		400:100	0.697	0.978	0.721	0.843	0.587	0.999
	Mixed	250:250	0.959	0.983	0.949	0.994	0.936	0.998
		100:400	0.737	0.844	0.831	0.893	0.615	0.952
		400:100	0.832	0.898	0.790	0.968	0.888	0.942

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

Latent Mean Difference of 0.5

Using a cutoff value of 0.05 in the equal factor loading conditions. Table 28 presents the model rejection rates of the SRMR when using a cutoff value of 0.05 in conditions where the true latent mean difference was equal to 0.5.

In the equal factor loading conditions, the incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were found to be high (in the range of 8.5% to 11.8%). When the factor-variance scaling method was implemented, the incorrect model rejection rates in the equal factor variance conditions, which can also be interpreted similarly to Type I error rates, were overly high (in the range of 18.7% to 20.8%). The correct model rejection rates in the unequal factor variance conditions, which can be interpreted similarly to statistical power, were found to be low (ranging from 32.9% to 50.1%).

Using a cutoff value of 0.05 in the unequal factor loading conditions. When using a cutoff of 0.05 in the unequal factor loading conditions, model rejection rates based on the RI strategy were generally higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.4, model rejection rates based on the RI strategy were all above 15% and several of them exceeded 90%. Within conditions with a loading difference of 0.1, most of the model rejection rates were between 10% and 14%. Average model rejection rates were 11.3% and 50.8% for the respective loading difference conditions of 0.1 and 0.4. Across the three factor variance ratio conditions, model rejection rates based on the RI strategy did not show substantial or systematic differences. Average model rejection rates were 31.6%, 33.2% and 28.2% for the 1:1, 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, respectively. Across the four factor loading patterns, different

trends were observed in the 0.1 and 0.4 loading difference conditions. Within conditions with a loading difference of 0.1, model rejection rates in the "mixed" pattern conditions(with an average rejection rate of 12.7%) were slightly higher than those in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions (with average rejection rates of 10.8%, 10.4% and 11%, respectively). When increasing the loading difference to 0.4, model rejection rates showed larger differences across the four factor loading patterns. More specifically, the "mixed" pattern generally led to higher model rejection rates than did the "1st loading unequal," "2nd loading unequal" and "all lower" patterns. In the "mixed" pattern conditions, all model rejection rates were above 60% and three of them were greater than 90%. Model rejection rates in the "all lower" pattern conditions were lower than those in the "mixed" pattern conditions but were higher than those in the "1st loading unequal" and "2nd loading unequal" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 34.4%, 33.3% 51.6% and 83.7% respectively for the "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions. Across the three sample size ratio conditions, model rejection rates in the 1:1 and 4:1 sample size ratio conditions were generally higher than those in the 1:4 sample size ratio conditions. In addition, model rejection rates were similar in the 1:1 and 4:1 sample size ratio conditions. Average model rejection rates were 35.4%, 24.1% and 33.6% for the 1:1, 1:4 and 4:1 sample size ratio conditions, respectively.

Table 28 also contains the model rejection rates based on the factor-variance scaling method when using a cutoff of 0.05 in conditions in which the true latent mean difference was equal to 0.5. Model rejection rates based on the factor-variance scaling method were higher than those based on the RI strategy. For the two loading difference magnitudes, model rejection rates based on the factor-variance scaling method were higher when the loading difference magnitude

was larger. Within conditions with a loading difference of 0.1, most of the model rejection rates were between 20% and 50%. Within conditions with a loading difference of 0.4, more than half of the model rejection rates were above 90%. Average model rejection rates were 39.8% and 82.9% for the loading difference conditions of 0.1 and 0.4, respectively. Across the three factor variance ratio conditions, model rejection rates based on the factor-variance scaling method were higher in the 0.8:1.2 factor variance ratio conditions than in the 1:1 and 1.2:0.8 factor variance ratio conditions; half of the model rejection rates under the 0.8:1.2 factor variance ratio conditions were greater than 90% and one of them was equal to 100%. For the factor variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, average model rejection rates were 53.2%, 54.1% and 76.6%, respectively. Regarding the model rejection rates across the four factor loading patterns, trends varied while the loading difference varied. When the true loading difference was set to 0.1, model rejection rates did not differ substantially or systematically as a function of the factor loading patterns. When the true loading difference was set to 0.4, model rejection rates were generally higher in the "all lower" and "mixed" pattern conditions (with average rejection rates of 93.8% and 94.1%, respectively) than in the "1st loading unequal" and "2nd loading unequal" pattern conditions (with average rejection rates of 72.2% and 71.3%, respectively). For the three sample size ratios, the 1:1 sample size ratio conditions generally led to higher model rejection rates than did the 1:4 and 4:1 sample size ratio conditions. Additionally, model rejection rates were similar in the 1:4 and 4:1 sample size ratio conditions. Average model rejection rates were 66.9%, 58.4% and 58.7% for the sample size ratio conditions of 1:1, 1:4 and 4:1, respectively.

Using a cutoff value of 0.08 in the equal factor loading conditions. Table 29 contains the model rejection rates when using a cutoff value of 0.08 in conditions where the true latent

mean difference was equal to 0.5. Model rejection rates obtained when using a cutoff value of 0.08 were much lower than those in Table 28 in which a cutoff value of 0.05 was used.

Table 29

Model Rejection Rates of the SRMR When Using a Cutoff of 0.08 in Conditions where the Latent Mean Difference is 0.5

			Factor Variance Ratio					
			1:1		1.2 :0.8		0.8:1.2	
Loading Difference	Loading Pattern	Sample Size Ratio	RI	FV	RI	FV	RI	FV
0	Equal Loading	250:250	*0.000*	*0.000*	*0.000*	0.001	*0.000*	0.000
		100:400	*0.000*	*0.000*	*0.000*	0.001	*0.000*	0.000
		400:100	*0.000*	*0.000*	*0.000*	0.002	*0.000*	0.002
0.1	1st Loading Unequal	250:250	0.000	0.000	0.000	0.000	0.000	0.003
		100:400	0.000	0.000	0.000	0.000	0.000	0.004
		400:100	0.000	0.000	0.000	0.001	0.000	0.001
	2nd Loading Unequal	250:250	0.000	0.000	0.000	0.001	0.000	0.008
		100:400	0.000	0.000	0.000	0.000	0.000	0.004
		400:100	0.000	0.000	0.000	0.000	0.000	0.005
	All Lower	250:250	0.000	0.000	0.000	0.001	0.000	0.035
		100:400	0.000	0.000	0.000	0.001	0.000	0.017
		400:100	0.000	0.000	0.000	0.000	0.000	0.009
	Mixed	250:250	0.000	0.000	0.000	0.004	0.000	0.005
		100:400	0.000	0.000	0.000	0.000	0.000	0.007
		400:100	0.000	0.001	0.000	0.001	0.000	0.000
0.4	1st Loading Unequal	250:250	0.000	0.019	0.000	0.001	0.000	0.367
		100:400	0.000	0.015	0.000	0.000	0.000	0.217
		400:100	0.000	0.003	0.000	0.000	0.000	0.052
	2nd Loading Unequal	250:250	0.000	0.017	0.000	0.001	0.000	0.333
		100:400	0.000	0.017	0.000	0.000	0.000	0.203
		400:100	0.000	0.002	0.000	0.000	0.000	0.059
	All Lower	250:250	0.000	0.529	0.004	0.051	0.001	0.971
		100:400	0.000	0.270	0.000	0.010	0.000	0.811
		400:100	0.002	0.157	0.001	0.005	0.001	0.670
	Mixed	250:250	0.032	0.062	0.020	0.234	0.019	0.250
		100:400	0.004	0.025	0.003	0.022	0.002	0.169
		400:100	0.006	0.028	0.003	0.207	0.013	0.036

Note. Model rejection rates that can be interpreted similarly to Type I error rates are italicized. The rest of the model rejection rates can be interpreted similarly to statistical power. Abbreviations used in this table are explained in Table 6.

In the equal factor loading conditions, the incorrect model rejection rates based on the RI strategy, which can be interpreted similarly to Type I error rates, were all equal to zero. When implementing the factor-variance scaling method, all three incorrect model rejection rates under the equal factor variance conditions, which can be interpreted similarly to Type I error rates, were all equal to zero. In addition, two out of six correct model rejection rates in the 1.2:0.8 and 0.8:1.2 factor variance ratio conditions, which can be interpreted similarly to statistical power, were equal to zero. The non-zero rejection rates were also very low (in the range of 0.1% to 0.2%).

Using a cutoff value of 0.08 in the unequal factor loading conditions. In the unequal factor loading conditions, model rejection rates based on the RI strategy were all equal to zero in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.4, model rejection rates based on the RI strategy were also equal to zero in the "1st loading unequal" and "2nd loading unequal" pattern conditions and increased slightly in the "all lower" and "mixed" pattern conditions. Average model rejection rates were 0% and 0.3% for the loading difference conditions of 0.1 and 0.4, respectively. Additionally, model rejection rates based on the RI strategy did not differ substantially or systematically as a function of the factor variance ratios or sample size ratios. Regarding the model rejection rates across the four factor loading patterns, clear trends were only found in conditions in which the loading difference was 0.4. More specifically, the model rejection rates in the "mixed" pattern conditions were generally higher than those in the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions. In the "mixed" pattern conditions, the model rejection rates were all greater than zero. In contrast, in the "1st loading unequal" and "2nd loading unequal" pattern conditions, all model rejection

rates were equal to zero. In the "all lower" pattern conditions, four out of nine model rejection rates were equal to zero.

Table 29 also contains the model rejection rates based on the factor-variance scaling method when using a cutoff value of 0.08 in conditions where the true latent mean difference was equal to 0.5. These model rejection rates were also lower than their counterparts in Table 28 in which a cutoff of 0.05 was used. The trends across loading difference magnitudes and factor variance ratios were also consistent with those observed when using a cutoff of 0.05. For example, model rejection rates based on the factor-variance scaling method were generally higher in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Within conditions with a loading difference of 0.1, all but one model rejection rates were equal to zero in the 1:1 factor variance ratio conditions. Half of the model rejection rates were equal to zero in the factor variance ratio conditions of 1.2:0.8. There was one rejection rate that was equal to zero in the factor variance ratio conditions of 0.8:1.2. Within conditions with a loading difference of 0.4, only four model rejection rates were equal to zero and they were all found in the 1.2:0.8 factor variance ratio conditions. The average rejection rates were 0.3% and 16.1% for the loading difference conditions of 0.1 and 0.4, respectively. In addition, model rejection rates based on the factor-variance scaling method were higher in conditions in which the factor variance ratio was 0.8:1.2 than in conditions in which the factor variance ratio was 1:1 or 1.2:0.8. In the 0.8:1.2 factor variance ratio conditions, only one model rejection rate was equal to zero, and some high and relatively high rejection rates (e.g., 97.1% and 81.1%) were observed. In the 1:1 and 1.2:0.8 factor variance ratio conditions, about half of the model rejection rates were equal to zero and most of the non-zero rejection rates were lower than 25%. For the factor

variance ratio conditions of 1:1, 1.2:0.8 and 0.8:1.2, average model rejection rates were 4.8%, 2.3% and 17.7%, respectively.

When using a cutoff of 0.08, the trends across the four loading patterns were slightly different from those observed when using a cutoff of 0.05. Specifically, within conditions with a loading difference of 0.4, model rejection rates were higher in the "all lower" pattern conditions than in the "the 1st loading unequal" and "2nd loading unequal" pattern conditions. Three largest model rejection rates (i.e., 97.1%, 81.1% and 67%) occurred in the "all lower" pattern conditions. In addition, the "all lower" pattern produced higher model rejection rates than did the "mixed" pattern in conditions in which the factor variance ratio was 1:1 or 0.8:1.2. When using a cutoff of 0.05, model rejection rates were similar in the "mixed" and "all lower" pattern conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 7.5%, 7.0%, 38.6% and 11.5% for the respective "1st loading unequal," "2nd loading unequal," "all lower" and "mixed" pattern conditions. Last, the trends across the three sample size ratios were different in the 0.1 and 0.4 loading difference conditions. In the 0.1 loading difference conditions, model rejection rates did not differ substantially or systematically as a function of the sample size ratios. In the 0.4 loading difference conditions, model rejection rates were generally higher in the equal sample size conditions than in the unequal sample size conditions. For the two unequal sample size conditions, the model rejection rates were slightly higher in the 1:4 sample size ratio conditions than in the 4:1 sample size ratio conditions. This trend was different from the trend observed when using a cutoff value of 0.05 in which model rejection rates did not differ systematically in the 1:4 and 4:1 sample size ratio conditions. Within conditions with a loading difference of 0.4, average model rejection rates were 23.6%, 14.7% and 10.2% for the sample size ratio conditions of 1:1, 1:4 and 4:1, respectively.

Chapter 5: Discussion

The primary question addressed in the present study was whether violating the assumptions underlying the RI strategy and/or the factor-variance scaling method (i.e., using a RI with non-invariant factor loadings or constraining unequal factor variances to a value of one across groups) would affect the testing and description of the latent mean difference across groups. The likelihood ratio test (LRT_{κ}), which has been used to test the significance of the latent mean difference across groups, was evaluated by assessing its Type I error rates and power under varying conditions. The standardized latent mean difference effect size measure ($\hat{\delta}_{\kappa}$), which has been proposed to describe the practical difference between two groups' latent means, was investigated by assessing its relative parameter bias and parameter bias under specified conditions. Additionally, the present study also examined the performance of model fit indices, including the χ^2 test of model fit, RMSEA, SRMR, CFI and TLI, with respect to correctly and incorrectly rejecting the null hypothesis of model fit. In this chapter, the results are discussed in the same order as they were presented in the previous chapter. First, Type I error rates associated with the LRT_{κ} are discussed, followed by the power results associated with the LRT_{κ}. Next, parameter bias and relative parameter bias of the $\hat{\delta}_{\kappa}$ are considered. The succeeding section examines model rejection rates of the χ^2 test, CFI, TLI, RMSEA and SRMR. After discussing the findings, implications and recommendations for applied researchers are provided. Finally, the limitations of the current study as well as additional topics for future research are provided.

Type I Error Rates of the LRT_{κ}

In the current study, when implementing the RI strategy, the Type I error rate of the LRT_{κ} was not adversely affected by loading difference magnitude, factor loading pattern, sample size ratio or factor variance ratio. That is, all of the observed Type I error rates when using the RI strategy were within Bradley's (1978) criterion of 0.05 ± 0.025. One result that is worthy of mentioning is that Type I error rates of the LRT_{κ} did not differ substantially from 0.05 when incorrectly constraining the first factor loadings to be equal across groups. This result indicated that violating the assumption of equivalent reference indicator loadings underlying the RI strategy did not affect Type I error rates associated with the LRT_{κ}. Among the four factors manipulated in this study, results regarding sample size ratio and factor loading pattern were consistent with the findings of previous research. Specifically, Hancock et al. (2000) found that Type I error rates of the LRT_{κ} were well controlled in both full and partial metric invariance conditions, regardless of how the sample size ratio varied.

Previous studies have not investigated Type I error rates of the LRT_{κ} when implementing the factor-variance scaling method. The findings in the current study indicated that when using the factor-variance scaling method, factor variance ratio, sample size ratio and loading difference magnitude affected Type I error rates associated with the LRT_{κ}. More specifically, in the unequal factor loading conditions, all Type I error rates that were beyond the criterion of 0.05 ± 0.025 occurred under the unequal sample size ratio conditions of 1:4 and 4:1. Additionally, most of the Type I error rates that were beyond the cutoff criterion were found in conditions in which the factor variance ratio was 0.8:1.2 and the loading difference was 0.4. Thus, when sample sizes and factor loadings were unequal for the two groups and the loading difference was relatively large (e.g., 0.4), constraining unequal factor variances (with a factor variance ratio of 0.8:1.2) to

a value of one across groups tended to lead to overly conservative or overly liberal Type I error rates. However, when the sample sizes were equal across groups, violating the equal factor-variance assumption (with a factor variance ratio of 1.2:0.8 or 0.8:1.2) did not have any substantial impact on Type I error rates of the LRT_{κ}.

Power of the LRT_{κ}

In this simulation study, it was found that power associated with the LRT_{κ} was affected by sample size ratio and loading difference magnitude. For example, in the equal factor loading conditions, power rates that were below the criterion of 0.90 (Goodman & Berlin, 1994) were found only in the unequal sample size conditions (with a sample size ratio of 1:4 or 4:1). This finding is consistent with previous research conducted by Kaplan and George (1995) who found that power was low when factor loadings were invariant across groups in unequal sample size conditions. In the unequal factor loading conditions manipulated in the current study, sample size ratio also influenced the power of the LRT_{κ}. More specifically, power rates fell below the criterion of 0.90 in seven unequal sample size ratio (1:4 or 4:1) conditions, five in which the RI strategy was implemented and two in which the factor-variance scaling method was implemented.

Loading difference magnitude was another factor that affected the power of the LRT_{κ}. First, in the unequal factor loading conditions, power rates lower than 0.90 universally occurred in conditions in which the loading difference was 0.1. Second, power increased slightly when the loading difference magnitude increased from 0.1 to 0.4. The increase in power can be explained in the context of construct reliability (see Equation 33). As mentioned before, construct reliability is influenced by the magnitude of the standardized factor loadings and the number of

loadings per factor in a CFA model. As the magnitude of the standardized factor loadings and the number of loadings per factor increase, construct reliability also tends to increase (Hancock, 2001). In the current study, while holding the model size and sample size ratio constant, increasing the loading difference magnitude resulted in increased factor loadings and, thus, increased construct reliability. According to Yang (2008), higher construct reliability leads to higher power in detecting latent mean differences across groups. Accordingly, the power associated with the LRT_{κ} was slightly higher in the 0.4 loading difference conditions than in the 0.1 conditions examined in this study.

One thing that should be noted is that although some power rates were found to be lower than the 0.90 criterion, they were in the range of 0.85 to 0.90. Most of the power rates in the current study were above 0.90 and several of them were equal to 1.00. High power was particularly observed in the large latent mean difference (0.5) conditions. Hancock et al. (2000) and Yang (2009) found that sufficient power with respect to detecting latent mean differences across groups can be achieved when the latent mean difference was equal to 0.50, regardless of other varying conditions.

Parameter Bias of the $\widehat{\delta}_{\kappa}$

Previous studies have not investigated parameter bias of the $\hat{\delta}_{\kappa}$ under varying conditions, particularly when the assumptions underlying the RI strategy and/or the factor-variance scaling method are violated. In the current study, all absolute values of parameter bias were below the cutoff value of 0.05, indicating that violating the assumptions associated with the RI strategy and/or the factor-variance scaling method did not have any substantial or systematic impact on the parameter bias of the $\hat{\delta}_{\kappa}$. Results also indicated that the parameter bias of the $\hat{\delta}_{\kappa}$ was not

affected by loading difference magnitude, sample size ratio, factor variance ratio or factor loading pattern.

Relative Parameter Bias of the $\widehat{\delta}_{\kappa}$

Similar to parameter bias of the $\hat{\delta}_{\kappa}$, relative parameter bias of the $\hat{\delta}_{\kappa}$ has not been examined in previous simulation research. In the present study, all relative parameter bias in the equal factor loading conditions was acceptable, indicating that violating the equal factor-variance assumption or varying sample size ratios did not affect the relative parameter bias of the $\hat{\delta}_{\kappa}$.

In the unequal factor loading conditions, however, it was found that loading difference magnitude, factor loading pattern and sample size ratio influenced the relative parameter bias of the $\hat{\delta}_{\kappa}$, regardless of the factor scaling method used. First, when increasing the magnitude of the loading difference from 0.1 to 0.4, the relative parameter bias of the $\hat{\delta}_{\kappa}$ also increased. More unacceptable relative parameter bias results were found in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. The increasing relative parameter bias in the larger loading difference conditions was anticipated. In the unequal factor loading conditions, the estimating models were incorrectly specified by constraining unequal factor loadings to be equal across groups. As a consequence, the equal factor loadings for the two groups were rescaled based on the constrained unequal factor loading values. When comparing latent means across groups, latent mean estimates for the two groups were readjusted due to the incorrect factor loading constraints across groups (see Equation 25). In the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions, factor loadings were consistently higher in the second group than in the first groups with the specified loading difference. Thus, latent mean estimates for the two groups were readjusted in opposite directions. When increasing the loading

difference magnitude, latent mean estimates for the two groups showed larger deviations from their actual values because the increasingly unequal factor loadings, when constrained to be equal, imposed more of a readjustment onto the latent mean estimates within each group in order to satisfy Equation 25. Consequently, latent mean difference estimates were less accurate and relative parameter bias increased.

Factor loading pattern also influenced the relative parameter bias of the $\hat{\delta}_{\kappa}$. More specifically, the "mixed" pattern led to more acceptable relative parameter bias than did the "1st loading unequal," "2nd loading unequal" and "all lower" patterns. This trend was anticipated because in the "mixed" pattern conditions, the magnitude of the factor loadings was the same for the two groups (only the position of the higher factor loading was different). When incorrectly constraining all factor loadings to be equal across groups, factor loadings and, thus, latent mean estimates for the two groups were readjusted in the same direction. Thus, latent mean differences between the two groups were accurately estimated, and more acceptable relative parameter bias was found than in the other three factor loading pattern conditions. Although relative parameter bias of the $\hat{\delta}_{\kappa}$ has not been previously investigated when varying factor loading patterns, researchers have examined the impact of factor loading patterns on the accuracy of measurement invariance tests. For example, Meade and Lautenschlager (2004) found that the power of the omnibus covariance invariance test and of the invariance test for a specific loading was higher in the "mixed" pattern conditions than in the "all lower" pattern conditions. In the "mixed" pattern conditions, the accuracy of parameter estimates was equivalent in the two groups. Thus, covariance invariance tests and a specific loading's invariance test, which depended upon the accuracy of parameter estimation in each group, were more accurate in detecting non-invariance.

In contrast, in the "all lower" pattern conditions, parameters were less accurately estimated in one group, which led to less accurate measurement invariance tests.

Sample size ratio was another factor that affected the relative parameter bias associated with the $\hat{\delta}_{\kappa}$. More acceptable relative parameter bias associated with the $\hat{\delta}_{\kappa}$ was found in the 1:4 sample size ratio conditions than in the equal 1:1 and the unequal 4:1 sample size ratio conditions. This result can be explained when considering sample size and construct reliability for each group. In the "1st loading unequal," "2nd loading unequal" and "all lower" pattern conditions, factor loadings were generated to favor the second group. That is, in conditions in which the sample size ratio was 1:4, the second group, which had the higher factor loading(s) (and therefore higher construct reliability), was also associated with the larger sample size. According to Gagné and Hancock (2006), larger sample size and higher construct reliability lead to more accurate parameter estimations. In the current study, relative parameter bias of the $\hat{\delta}_{\kappa}$ was lower in the 1:4 sample size ratio conditions than in the 1:1 or 4:1 sample size ratio conditions because in the 1:4 sample size ratio conditions, the larger sample size and higher construct reliability in the second group, which led to more accurate parameter estimations, compensated for the less accurate parameter estimations in the first group.

It is important to note that although loading difference magnitude, factor loading pattern and sample size ratio affected the relative parameter bias of the $\hat{\delta}_{\kappa}$, an Analysis of Variance (ANOVA) using these conditions as independent variables of relative parameter bias outcomes in each replication indicated that none of the main effects or interaction effects were associated with effect size values (partial eta squared , η_p^2, values) greater than 0.06.

Model Rejection Rates Associated with the χ^2 Test of Model Fit

Model rejection rates associated with the χ^2 test were investigated under two latent mean difference magnitudes (0.0 and 0.5). This section first discusses the trends in conditions where the true latent mean difference was equal to zero, followed by the trends in conditions where the true latent mean difference was equal to 0.5.

Latent Mean Difference of Zero

In the equal factor loading conditions, model rejection rates when the RI strategy was used are actually equal to Type I error rates since the estimating models were correctly specified. Model rejection rates when the factor-variance scaling method was used are also equal to Type I error rates in the equal factor variance conditions, but are equal to statistical power in the unequal factor variance conditions since factor variances were incorrectly constrained to a value of one across groups. In the equal factor loading conditions, results indicated that neither Type I error rates when the RI strategy was used nor Type I error rates and power when the factor-variance scaling method was used were adversely affected by sample size ratio or factor variance ratio.

In the unequal factor loading conditions, model rejection rates of the χ^2 test can be interpreted as statistical power since the estimating models were incorrectly specified by constraining all factor loadings to be equal across groups. Results indicated that when implementing the RI strategy, the power of the χ^2 test was influenced by loading difference magnitude, factor loading pattern and sample size ratio. More specifically, when the loading difference increased from 0.1 to 0.4, power also increased. Higher power was observed in the 0.4 loading difference conditions because larger factor loading discrepancies were constrained to be

equal across the two groups, yielding more model misspecification. Thus, the χ^2 test tended to correctly reject the null hypothesis of model fit more frequently.

Factor loading pattern also affected the power of the χ^2 test of model fit. The impact of factor loading pattern was more obvious in the 0.4 loading difference conditions than in the 0.1 loading difference conditions. Specifically, within conditions with a loading difference of 0.4, the "mixed" pattern led to higher model rejection rates than did the "1st loading unequal," "2nd loading unequal" and "all lower" patterns. Sample size ratio, like loading difference magnitude and factor loading pattern, also influenced the power of the χ^2 test. To be precise, power of the χ^2 test was generally higher in the equal sample size conditions than in the unequal sample size conditions. When the sample sizes were unequal, power was generally higher in the 4:1 sample size ratio conditions than in the 1:4 sample size ratio conditions.

When implementing the factor-variance scaling method, the power of the χ^2 test was generally higher than those observed when using the RI strategy. This result was anticipated because using the factor-variance scaling method resulted in a less parameterized model than did using the RI strategy. Thus, χ^2 statistics (and therefore model rejection rates) were higher when using the factor-variance scaling method than when using the RI strategy. When implementing the factor-variance scaling method, it was found that loading difference magnitude, factor loading pattern, sample size ratio and factor variance ratio had an impact on the power of the χ^2 test. The trends across loading difference magnitudes and sample size ratios were consistent with those observed when using the RI strategy. However, the trends across factor loading patterns were slightly different from those observed when using the RI strategy. For instance, within conditions with a loading difference of 0.4, power tended to be higher in the "mixed" and the "all lower" pattern conditions than in the "1st loading unequal" and "2nd loading unequal" pattern

conditions. This result indicated that when increasing the percent of non-invariant factor loadings, the power associated with the χ^2 test also increased. This result, too, was expected because increasing the percent of non-invariant factor loadings led to larger model misspecification when factor loadings were constrained to be invariant across the two groups. Consequently, the χ^2 test tended to reject the null hypothesis of model fit more frequently than in conditions with a lower percent of non-invariant factor loadings. Additionally, when implementing the factor-variance scaling method, factor variance ratio had an impact on the power of the χ^2 test. In conditions in which the factor variance ratio was 0.8:1.2, the power of the χ^2 test was higher than in conditions in which the factor variance ratio was 1:1 or 1.2:0.8.

Latent Mean Difference of 0.5

In the equal factor loading conditions, Type I error rates and the power associated with the χ^2 test in conditions where the true latent mean difference was equal to 0.5 were consistent with those in conditions where the true latent mean difference was equal to zero. In the unequal factor loading conditions, the trends in power across loading difference magnitudes, factor loading patterns and factor variance ratios were also consistent with those in conditions where the true latent mean difference was equal to zero. Only the power trends across the three sample size ratios were slightly different from those found in the zero latent mean difference conditions. Specifically, when implementing the RI strategy, power in the equal sample size conditions was not consistently higher than those in the 4:1 sample size ratio conditions. Instead, the power in these two conditions did not differ substantially or systematically.

Model Rejection Rates of the CFI and TLI

The performance of the CFI and TLI with respect to correctly and incorrectly rejecting the null hypothesis of model fit was assessed using cutoff values of 0.90 and 0.95, respectively, in conditions where the true latent mean difference was equal to zero or 0.5.

Latent Mean Difference of Zero

In the equal factor loading conditions, incorrect model rejection rates when the RI strategy was implemented can be interpreted similarly to Type I error rates. Model rejection rates when the factor-variance scaling method was used can also be interpreted similarly to Type I error rates in the equal factor variance conditions, and can be interpreted similarly to statistical power in the unequal factor variance conditions. When using a CFI and TLI cutoff value of 0.90, a few incorrect model rejection rates when using the RI strategy were found to differ substantially from 5%, particularly in the unequal sample size conditions. This result indicated that when using a cutoff value of 0.90, the CFI and TLI tended to incorrectly reject model fit more often in the unequal sample size conditions than in the equal sample size conditions. The correct model rejection rates when using the factor-variance scaling method were not influenced by factor variance ratio or sample size ratio.

In the unequal factor loading conditions, model rejection rates as determined by the CFI and TLI can be interpreted similarly to statistical power. Using a cutoff value of 0.90, it was found that loading difference magnitude, factor loading pattern and sample size ratio affected the correct model rejection rates when the RI strategy was implemented. Loading difference magnitude, factor loading pattern, sample size ratio and factor variance ratio all had an impact on the correct model rejection rates when the factor-variance scaling method was implemented. All

the trends were consistent with those observed when investigating the χ^2 test of model fit in conditions where the true latent mean difference was equal to 0.5.

When using a CFI and TLI cutoff value of 0.95, model rejection rates were generally higher than those observed when using a cutoff value of 0.90, regardless of the factor scaling method used. For example, in the equal factor loading conditions, all of the incorrect model rejection rates were much higher than 5% when using either of the factor scaling methods. The correct model rejection rates when using the factor-variance scaling method were also higher than those found when using a cutoff value of 0.90, although they were still lower than the power criterion of 0.90. When using a cutoff value of 0.95 in the unequal factor loading conditions, the correct model rejection rates of the CFI and TLI also increased and demonstrated similar trends as those found when using a cutoff value of 0.90. These results indicated that violating the assumptions underlying the RI strategy and/or the factor-variance scaling method did not influence the model rejection rates as determined by the CFI and TLI. In addition, using the more stringent cutoff value (0.95) led to higher incorrect and correct model rejection rates of the CFI and TLI than did using the less stringent cutoff value (0.90).

Latent Mean Difference of 0.5

When the true latent mean difference was equal to 0.5, the model rejection rates of the CFI and TLI in the equal factor loading conditions were similar to the model rejection rates found in the corresponding zero latent mean difference conditions, regardless of cutoff criterion value used (0.90 or 0.95). Model rejection rates in the unequal factor loading conditions, however, were slightly higher than those in the corresponding zero latent mean difference conditions. Nonetheless, the model rejection rates demonstrated similar trends as those found

when the true latent mean difference was equal to zero. Further, the more stringent cutoff criterion value (0.95) led to higher model rejection rates than did the less stringent cutoff criterion value (0.90).

Model Rejection Rates of the RMSEA and SRMR

In this simulation study, the performance of the RMSEA and SRMR in terms of correct and incorrect model rejection rates was examined in conditions where the true latent mean difference was equal to zero or 0.5. Two RMSEA cutoff values (0.05 and 0.06) and two SRMR cutoff values (0.05 and 0.08) were used, respectively, to determine whether the null hypothesis of model fit should be rejected.

Latent Mean Difference of Zero

When using the more stringent RMSEA and SRMR cutoff value of 0.05 in the equal factor loading conditions, some of the incorrect model rejection rates when using the RI strategy and a few incorrect model rejection rates when using the factor-variance scaling method were found to differ substantially from 5%. The difference between the two model fit indices was that the incorrect model rejection rates of the RMSEA were lower than 5% whereas the incorrect model rejection rates of the SRMR were much higher than 5%. Although the correct model rejection rates of the RMSEA and the SRMR when using the factor-variance scaling method were lower than 90%, correct model rejection rates of the SRMR were much higher than those of the RMSEA. Results also indicated that in the equal factor loading conditions, model rejection rates associated with the RMSEA and SRMR did not vary as a function of sample size ratio or factor variance ratio.

When using the RMSEA and SRMR cutoff value of 0.05 in the unequal factor loading conditions, the model rejection rate trends across loading difference magnitude, factor loading pattern and factor variance ratio were consistent with those found when using the CFI and TLI. However, the trends across sample size ratios were slightly different from those associated with the CFI and TLI. To elaborate, when the loading difference was large (0.4), model rejection rates associated with the RMSEA and SRMR were generally higher in the equal 1:1 sample size ratio conditions than in the unequal 1:4 and 4:1 conditions whereas model rejection rates associated with the CFI and TLI in the 1:1 and 4:1 sample size ratio conditions were similar but were higher than those in the 1:4 sample size ratio conditions.

When using the less stringent RMSEA cutoff value of 0.06 and the less stringent SRMR cutoff value of 0.08, model rejection rates dropped greatly in both equal and unequal factor loading conditions. This trend was more prevalent when using the SRMR since most of the model rejection rates of the SRMR were equal to or close to zero when using a cutoff value of 0.08. In addition, the rejection rate trends were consistent with those observed when using the more strict RMSEA and SRMR cutoff value of 0.05.

Latent Mean Difference of 0.5

In conditions with a true latent mean difference equal to 0.5, model rejection rates associated with the RMSEA and SRMR were slightly higher than those observed in the corresponding conditions in which the true latent mean difference was equal to zero. This trend was observed regardless of using the more stringent or less stringent cutoff values. The rejection rate trends across loading difference magnitude, factor loading pattern, factor variance ratio and sample size ratio conditions when the true latent mean difference was equal to 0.5 were all

consistent with trends found in corresponding conditions where the true latent mean difference was equal to zero. Using the more stringent cutoff value also led to higher model rejection rates associated with the RMSEA and SRMR, which was consistent with the trend observed in conditions where the true latent mean difference was equal to zero.

In sum, the results of the current study indicated that when the RI strategy was implemented, factor loading magnitude, factor loading pattern and sample size ratio affected model rejection rates of the χ^2 test of model fit, CFI, TLI, RMSEA and SRMR. When using the factor-variance scaling method, factor loading magnitude, factor loading pattern, sample size ratio and factor variance ratio influenced model rejection rates of the same model fit indices mentioned above. Results also demonstrated that using the more stringent cutoff values led to higher model rejection rates than did using the less stringent cutoff values. Further, when using the more stringent cutoff value, the SRMR performed better than the CFI, TLI and RMSEA with respect to correctly rejecting the null hypothesis of model fit. However, when using the less stringent cutoff value, the SRMR performed much worse as compared to the remaining three model fit indices in identifying model mis-specifications.

Implications and Recommendations

Based on the results of the current study, violating the assumption underlying the RI strategy (i.e., using a RI with non-invariant factor loadings across groups) did not affect the Type I error rates of the LRT_{κ}. However, violating the assumption associated with the factor-variance scaling method (i.e., constraining unequal factor variances to a value of one across groups) influenced the Type I error rates of the LRT_{κ}, particularly when sample sizes and factor loadings were unequal across groups. In addition, the power of the LRT_{κ} was not affected by violating the

assumptions underlying the RI strategy and/or the factor variance scaling method. Instead, sample size ratio and loading difference magnitude influenced the power of the LRT_{κ}. Regarding parameter bias and relative parameter bias of the $\hat{\delta}_{\kappa}$, results indicated that only relative parameter bias of the $\hat{\delta}_{\kappa}$ varied as a function of the loading difference magnitude, factor loading pattern and sample size ratio. Neither parameter bias nor relative parameter bias was affected by violating the assumptions underlying the RI strategy and/or the factor-variance scaling method. Results also demonstrated that the performance of model fit indices, including the χ^2 test of model fit, CFI, TLI, RMSEA and SRMR with respect to correct model rejection rates was influenced by loading difference magnitude, factor loading pattern, sample size ratio and factor variance ratio. However, violating the assumptions underlying the two factor scaling methods did not have any systematic impact on model rejection rates associated with the five model fit indices examined.

According to Johnson et al. (2009), violating the assumption underlying the RI strategy does not affect the accuracy of the full metric invariance test but has an impact on the accuracy of a specific loading's invariance test. The present study provides further information regarding the impact of violating the assumptions underlying the RI strategy and/or the factor variance scaling method on the latent mean difference test and effect size measure of the latent mean difference across groups. Specifically, researchers do not necessarily need to be concerned about violating the assumption underlying the RI strategy given that it does not affect the performance of the LRT_{κ}. This finding is valuable because the assumption underlying the RI strategy may be frequently violated since it is difficult to identify an item with truly invariant factor loadings to serve as a RI in practice. In contrast, researchers should be aware of the assumption underlying the factor-variance scaling method. In particular, when the sample sizes for the two groups being

compared are unequal, constraining unequal factor variances to a value of one across groups is likely to produce overly conservative or overly liberal Type I error rates associated with the LRT_{κ}. When using the $\hat{\delta}_{\kappa}$ in order to describe the latent mean difference across groups, researchers do not necessarily need to be concerned about violating the assumptions underlying the two factor scaling method given that parameter bias and relative parameter bias estimates of the $\hat{\delta}_{\kappa}$ were not adversely affected.

When using the χ^2 test, CFI, TLI, RMSEA and SRMR to evaluate model fit, more stringent cutoff values (e.g., 0.95 for the CFI and TLI and 0.05 for the RMSEA and SRMR) are recommended because they lead to more correct identification of model mis-specification. Another important finding is that model rejection rates associated with the χ^2 test, CFI, TLI, RMSEA and SRMR are not affected by violating the assumptions underlying the RI strategy and/or factor variance scaling method. However, they are influenced by the loading difference magnitude, sample size ratio, factor variance ratio and factor loading pattern. Thus, researchers should avoid using a single index to evaluate model fit because of the possible impact of these factors.

Limitations and Suggestions for Future Research

The assumptions underlying the RI strategy and the factor-variance scaling method have not been widely investigated in previous simulation studies. Thus, as a starting point for this line of research, the current study included a relatively simple model and investigated more ideal conditions. Due to the preliminary nature of the research, there are several limitations inherent in the present study. First, only a large latent mean difference was included when investigating the power of the LRT_{κ}. As a result, power associated with the LRT_{κ} was high in these conditions and

did not differ systematically as a function of the factor loading pattern or factor variance ratio. It was found that violating the assumptions underlying the two factor-scaling methods did not influence the power of the LRT_{κ}. However, it is not clear whether the same findings would be obtained with smaller latent mean differences (e.g., 0.1 and 0.3). In future simulation studies, researchers could include smaller latent mean differences and examine whether violating the assumptions underlying the two factor-scaling method would affect the power of the LRT_{κ}.

Second, only two moderately unequal factor variance conditions were included in the present study. In the equal factor loading conditions, it was found that correct model rejection rates associated with the χ^2 test, CFI, TLI, RMSEA and SRMR were low, regardless of the cutoff value used and latent mean difference magnitude. It was not clear whether the low correct model rejection rates were caused by incorrectly constraining unequal factor variances to a value of one across groups or were due to the nature of the model fit indices. Future studies could include more extreme factor variance ratio conditions to further investigate whether violating the equal factor-variance assumption would affect the model rejection rates of those indices. Also, when violating the assumption underlying the factor-variance scaling method with a factor variance ratio of 0.8:1.2 in the unequal factor loading conditions, the correct model rejection rates associated with the χ^2 test, CFI, TLI, RMSEA and SRMR were not adversely affected. Instead, correct model rejection rates were higher under the 0.8:1.2 factor variance ratio conditions than under the 1:1 and 1.2:0.8 factor variance ratio conditions. It is not clear why this result was observed given these limited conditions of factor variance ratio discrepancies. Thus, future studies could include more factor variance ratio conditions to explore the possible reasons for this finding.

Third, model size and model complexity were not varied in the current study. To simplify the design of this simulation study, a two-group, one-factor, six-indicator CFA model was used throughout. Future researchers could consider more complex models (e.g., more observed indicators, additional latent variables and three or more groups) to investigate whether varying the model size and/or model complexity would affect the testing and description of the latent mean difference across groups. In addition, researchers who include models with more observed indicators could likewise investigate more severe loading non-invariance conditions.

Finally, the generalized variance (as measured by the determinants of the covariance matrix) was not manipulated. Previous research has indicated that the generalized variance affects the power of the latent mean difference test. For example, Kaplan and George (1995) found that when a group with the larger generalized variance was associated with the larger sample size, the power of the latent mean difference test was higher than in conditions in which the group with the smaller generalized variance was paired with the larger sample size. Future researchers could manipulate factor loadings, factor variances and error variances for groups to create conditions in which the larger generalized variance is paired with the larger or smaller sample size.

General Conclusion

The RI strategy and factor-variance scaling method are two approaches that are commonly used to set the scale of the latent variable before comparing latent means across groups. This study indicates that violating the assumption underlying the RI strategy does not adversely affect the testing and description of the latent mean difference across groups. However, violating the assumption associated with the factor-variance scaling method influences

the Type I error rate of the likelihood ratio test. This study also demonstrates that model rejection rates of model fit indices, including the χ^2 test of model fit, CFI, TLI, RMSEA and SRMR, vary as a function of the factor loading magnitude, factor loading pattern, sample size ratio and factor variance ratio. In addition, when using the more stringent cutoff values, these model fit indices perform better in correctly identifying model misspecification. It is hoped that this study provides researchers with useful information concerning the performance of the LRT_{κ}, the $\hat{\delta}_{\kappa}$ and model fit indices under factor scaling method assumption violations.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155-173.

Babyak, M. A., Synder, C. R., & Yoshinobu, L. (1993). Psychometric properties of the Hope Scale: A confirmatory factor analysis. *Journal of Research in Personality, 27,* 154-169.

Bentler, P. M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures. *Psychometrika, 48,* 493-571.

Bentler, P. M. (1989). *EQS structural equations program manual*. Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Boomsma, A. (1983). *On the robustness of LISREL (Maximum Likelihood Estimation) against small sample size and nonnormality*. Amsterdam: Sociometric Research Foundation.

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (2008). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-III in the United States and Canada. *Educational and Psychological Measurement, 68*(6), 1024-1040.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Browne, M. W., & Mels, G. (1990). *RAMONA user's guide*. Unpublished report, Department of Psychology, The Ohio State University, Columbus.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and means structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypothesis: A review of contemporary of management research recommendations for future studies. *Organizational Research Methods, 7*(2), 151-167.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences and strategies. *Sociological Methods and Research, 29*(4), 468-508.

Cheung, G. W. & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

Chou, C. P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavior Research, 25,* 115-136.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 2,* 317-327.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46,* 137-149.

Enders, C. K., & Finney, S. J. (2003, April). *SEM fit index criteria re-examined: An investigation of ML and robust fit indices in complex models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Fan, X., & Fan, X. (2005). Using SAS for Monte Carlo simulation research in SEM. *Structural Equation Modeling, 12*(2), 299-333.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: locating the invariant referent sets. *Structural Equation Modeling, 15,* 96-113.

Gagné, P., & Furlow, C. F. (2009). Automating multiple software packages in simulation research for structural equation modeling and hierarchical linear modeling. *Structural Equation Modeling, 16*(1), 179-185.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41*(1), 65-83.

Gierl, M. J., & Mulvenon, S. (1995, April). *Evaluation of the application of fit indices to structural equation models in educational research: a review of literature from 1990 through 1994*. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: every "one" matters. *Psychological Methods, 6*(3), 258-269.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*(3), 200-206.

Hancock, R. G. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30,* 91-105.

Hancock, R. G. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66,* 373-388.

Hancock, R. G., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7(4),* 534-556.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967-988.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a Meta-Analysis. *Sociological Methods Research, 26,* 329-367.

Hu, L., & Bentler, P. M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1),* 1-55.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16,* 642-657.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IL: Scientific Software.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling, 2,* 101-118.

Kim, S., Beretvas, S. N., & Sherry, A. R. (2010). A validation of the factor structure of OQ-45 scores using factor mixture modeling. *Measurement and Evaluation in Counseling and Development, 42*(4), 275-295.

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance tests of creative thinking-Figural. *Educational and Psychological Measurement, 66*(3), 459-477.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Lawrence, F. R., & Hancock, G. R. (1998, April). *Finite sample behavior of the likelihood ratio, Wald, and Lagrange Multiplier tests: Bias and variability in univariate noncentrality parameter estimation.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Manolis, C., Levin, A., & Gahlstrom, R. (1997). A Generation X scale: Creation and Validation. *Educational and Psychological Measurement, 57,* 666-684.

Meade, A. W. & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*(1), 60-72.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97-103.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247-255.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*, 60-72.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153-172). Mahwah. NJ: Lawrence Erlbaum Associates, Inc.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73,* 574-584.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93-115.

Riordan, C. R., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20,* 643-671.

Rossi, J. S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Statistical Psychology, 58,* 646-656.

Sabiston, C. M., & Crocker, P. R. E. (2007). Exploring self-perceptions and social influences as correlates of adolescent leisure-time physical activity. *Journal of Sport and Exercise Psychology, 30,* 3-22.

Smith, C. S., Tisak, J., Bauman, T., & Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: implications for between-and within-population assessments. *Journal of Applied Psychology, 76,* 628-636.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78-90.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253-264.

Stevens, J. (2002). Applied multivariate statistics for the social sciences (4th ed.). Hillsdale, NJ: Erlbaum.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structural models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 42,* 233-239.

Thompson, M. S., & Green, S. B. (2006). Evaluation between-group differences in latent variable means. In Serlin, R. C. (Ed.), *Structural Equation Modeling: A Second Course* (pp. 119-170). Greenwich, Connecticut: IAP.

Tofighi, D. (2005). *The effect of partial scalar invariance on multiple group comparison: t test and latent mean z test* (Master's thesis). University of Nebraska-Lincoln.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Van de Vijver, F. J. R., & Harsveld, M. (1994).The incomplete equivalence of the paper-and-pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology, 79*, 852-859.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, and S. G. West (eds.), *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research* (pp. 281-234). Washington, D. C.: American Psychological Association.

Yang, Y. (2008). *Partial invariance in loadings and intercepts-their interplays and implications for latent mean comparisons* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No.3297716)

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435-463.

Yuan, K., & Bentler, P. M. (2004). On chi-square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*(5), 737-757.

Vita

Dandan Wang was born in Zhengzhou, Henan province, China on February 15, 1981, the daughter of Yan Wang and Hong Zhou. In 1999, Dandan entered GuangDong University of Foreign Studies, Guangzhou, China majoring in English. She received the degree of Bachelor of Arts in June 2003. From August 2003 to May 2005, she studied at the Graduate School of the University of Missouri-Kansas City, majoring in Curriculum and Instruction. She received a Master of Arts from the University of Missouri-Kansas City. In August 2005, Dandan entered the Graduate School of the University of Missouri-Columbia, majoring in Educational Psychology with a focus on Quantitative Methods. One year later, she transferred to the University of Texas at Austin with her adviser and continued her graduate study in Educational Psychology in the area of Quantitative Methods. During her graduate study at UT-Austin, she worked as a teaching assistant in the Department of Educational Psychology and Division of Statistic and Scientific Computation. In addition, she worked as a graduate research fellow in the Division of Statistic and Scientific Computation in 2010.

This dissertation was typed by the author.